American Psychological Association

Volume 109
Number 2
February 2017

Published eight times

ISSN 0022-0663

Journal of Educational Psychology

www.apa.org/pubs/journals/edu

The main purpose of the *Journal of Educational Psychology*® is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Single Issues, Back Issues, and Back Volumes: For information regarding single issues, back issues, or back volumes, write to Order Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242; call 202-336-5600 or 800-374-2721; or visit www.apa.org/pubs/journals/subscriptions.aspx

Manuscripts: Submit manuscripts electronically through the Manuscript Submissions Portal found at www.apa.org/pubs/journals/edu according to the Instructions to Authors found elsewhere in this issue (see table of contents). Correspondence regarding manuscripts should be sent to the Editor, Steve Graham, at steve.graham@asu.edu. The opinions and statements published are the responsibility of the authors, and such opinions and statements do not necessarily represent the policies of APA or the views of the Editor.

Electronic Access: APA members who subscribe to this journal have automatic access to all issues of the journal in the PsycARTICLES® full-text database. See http://my.apa.org/access.html.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

APA Journal Staff: Rosemarie Sokol-Chang, PhD, *Publisher, APA Journals*; Annie Hill, *Managing Director*; John Hill, *Journal Production Manager*; Amanda Conley, *Editorial Manuscript Coordinator*; Jodi Ashcraft, *Director, Advertising Sales and Exhibits.*

Journal of Educational Psychology® (ISSN 0022-0663) is published eight times (January, February, April, May, July, August, October, November) in one volume per year by the American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Subscriptions are available on a calendar year basis only (January through December). The 2017 rates follow: *Nonmember Individual:* $250 Domestic, $292 Foreign, $314 Air Mail. *Institutional:* $953 Domestic, $1,030 Foreign, $1,054 Air Mail. *APA Member:* $123. *APA Student Affiliate:* $75. Write to Subscriptions Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Printed in the U.S.A. Periodicals postage paid at Washington, DC, and at additional mailing offices. POSTMASTER: Send address changes to *Journal of Educational Psychology*, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

Effective with the 1986 volume, this journal is printed on acid-free paper.

Journal of

Educational Psychology®

www.apa.org/pubs/journals/edu

February 2017

Volume 109
Number 2

Pre-School and Kindergarten

Aggression in School

Learning and Achievement

Reading and Mathematics

(*Contents continue*)

Creativity

Other

Journal of Educational Psychology
2017, Vol. 109, No. 2, 149–165

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000137

Pre-K Classroom-Economic Composition and Children's Early Academic Development

Portia Miller and Elizabeth Votruba-Drzal
University of Pittsburgh

Meghan McQuiggan
American Institutes for Research, Washington, DC

Alyssa Shaw
National Abortion Federation, Washington, DC

There are currently 2 principal models of publicly funded prekindergarten programs (pre-K): targeted pre-K, which is means-tested, and universal pre-K. These programs often differ in terms of the economic characteristics of the preschoolers enrolled. Studies have documented links between individual achievement in school-age children and the economic composition of classroom peers, but little research has revealed whether these associations hold in pre-K classrooms. Using data from 2,966 children in 709 pre-K classrooms, we examined whether classroom-economic composition (i.e., average family income, standard deviation of incomes, and percentage of students from low-income households) relates to achievement in preschool. Furthermore, this study investigated whether associations between classroom-economic composition and achievement differed depending on initial academic skill level. Increased economic advantage in pre-K classrooms positively predicted spring achievement. Specifically, increasing aggregate classroom income between $22,500 and $62,500 was related to improvements in math scores. Increases in the proportion of children from low-income households in the classroom were negatively related to both math and literacy and language skills when increases occurred between 52.5% and 72.5% and 25% and 45%, respectively. There was limited evidence that links between classroom-economic composition and achievement differed depending on initial skill level. Results suggest that economically integrated pre-K programs may be more beneficial to preschoolers from low-income households' achievement than classrooms targeting economically disadvantaged children.

Keywords: prekindergarten, academic achievement, classroom-economic composition

Not all children begin kindergarten on equal footing academically. Disparities in children's academic skills at kindergarten entry related to family socioeconomic status (SES) are well documented (e.g., Duncan & Magnuson, 2011; Garcia, 2015). Increasingly, prekindergarten (pre-K)—defined here as publicly funded, center-based preschool programs attended 1–2 years before kindergarten—has been heralded as a policy lever to narrow early socioeconomic gaps in achievement (Magnuson & Waldfogel, 2005). Indeed, mounting evidence shows positive impacts of pre-K on early academic skills (Gormley, Gayer, Phillips, & Dawson, 2005; Henry et al., 2003; Magnuson, Meyers, Ruhm, & Waldfogel, 2004; Reynolds & Temple, 1995; Weiland & Yoshikawa, 2013). Studies have found overwhelmingly that the benefits of pre-K are stronger for socioeconomically disadvantaged children than their more advantaged peers (Magnuson et al., 2004; Magnuson, Ruhm, & Waldfogel, 2007; Votruba-Drzal, Coley, Koury, & Miller, 2013).

As government budgets become more constrained and demand for pre-K programs grows, a key question for early-childhood education researchers and policymakers is whether it is more effective to offer targeted pre-K programs that limit enrollment solely to economically disadvantaged children or to implement universal pre-K, which is available to all children regardless of family income. Head Start, for example, is the only federally funded preschool program in the United States and is targeted, serving nearly one million children from low-income households (Office of Head Start, 2013). In addition, 40 states fund pre-K programs, and although most of these programs serve disadvantaged children, who are most often defined as children with family incomes below 185%–200% of the federal poverty level (FPL), four states and the District of Columbia have adopted universal pre-K programs (Barnett, Carolan, Squires, Brown, & Horowitz, 2014). These varied approaches to pre-K may lead to differences in the concentration of children from low-income households in preschool classes, thereby raising important questions regarding the role of classroom-economic composition—the collective eco-

This article was published Online First June 23, 2016.

Portia Miller and Elizabeth Votruba-Drzal, Department of Psychology, University of Pittsburgh; Meghan McQuiggan, American Institutes for Research, Washington, DC; Alyssa Shaw, National Abortion Federation, Washington, DC.

We thank the National Center for Early Development and Learning investigators, especially Diane Early, for access to and assistance with the data. We also thank the children, families, and teachers that participated in the Multi-State and the SWEEP studies.

Correspondence concerning this article should be addressed to Portia Miller, Department of Psychology, University of Pittsburgh. 4617 Sennott Square, 210 South Bouquet Street, Pittsburgh, PA 15260. E-mail: Plm11@pitt.edu

nomic characteristics (e.g., family income, poverty status) of students in a classroom—in shaping the development of children's school readiness skills. Although many studies have documented positive relations between increased classroom socioeconomic advantage and individual children's achievement in elementary and secondary school (see van Ewijk & Sleegers, 2010), scant research has examined whether these associations hold in preschool classrooms and for whom they are strongest.

Analyzing data from 2,966 English- and/or Spanish-speaking children in 709 pre-K classrooms across 11 states, this study examined whether aggregate pre-K classroom-economic composition, including average family income, variability of classroom incomes, and proportion of children in the class that are low-income, is related to individual academic achievement across the pre-K year. Moreover, we considered whether associations between classroom-economic composition and achievement are linear or whether they vary across levels of aggregate economic advantage. Finally, we investigated whether links between classroom-economic composition and individual achievement differ for children based on their academic skills at the start of pre-K. This study seeks to inform programs and policies aimed at enhancing early education and school readiness by elucidating how the economic characteristics of children in a preschool classroom relate to individual children's learning.

Theoretical Framework for Peer-Economic Compositional Effects in Pre-K

Children's academic development during pre-K is driven by proximal processes and interactions in the classroom context (Creemers & Reezigt, 1996). Children both affect and are influenced by the classroom learning environment. Proximal processes and interactions within the classroom are also shaped by the collective characteristics of students within it, including classroom-economic composition. When thinking about pre-K classrooms in particular, classroom-economic composition likely affects children's learning in three ways: by influencing (a) peer interactions, (b) teachers' instruction practices and interactions with children, and (c) the classroom structure.

First, children may be directly impacted by peers in their classrooms through day-to-day interactions that provide opportunities for modeling and reinforcement of behaviors and skills (e.g., Bub, McCartney, & Willet, 2007; Raver et al., 2011; Schechter & Bye, 2007). Children from low-income homes tend to possess fewer academic skills and exhibit more behavioral problems than their more advantaged peers (Magnuson & Votruba-Drzal, 2009). Thus, economically integrated classrooms may expose children from low-income households to peers that model more advanced academic skills and adaptive approaches to learning and classroom behavior. Conversely, in more disadvantaged pre-K classrooms, larger numbers of disruptive and aggressive peers may model maladaptive behavior that may be reinforced by classmates (Battistich et al., 1995; Dishion, Spracklen, Andrews, & Patterson, 1996), which could interfere with learning by negatively impacting children's attention and classroom behavior (Georges, Brooks-Gunn, & Malone, 2012; Hinshaw, 1992; Neidell & Waldfogel, 2010).

Second, differences in classroom-economic composition may elicit different instruction from teachers. Disadvantaged children tend to receive less instructive and evaluative feedback and engage in fewer responsive and positive interactions with teachers (e.g., Arnold, 1997; Connor, Son, Hindman, & Morrison, 2005). Classrooms with high levels of disadvantage are often characterized by less constructivist, student-centered instruction (Stipek, 2004), which may inhibit achievement. In addition, if greater numbers of students in highly disadvantaged pre-Ks are struggling with stress at home, learning delays, or attention and behavior problems, teachers may spend more time tending to these needs instead of focusing on instruction (e.g., Carr, Taylor, & Robinson, 1991; Dreeben & Barr, 1988; Gamoran, 1986; Lavy, Paserman, & Schlosser, 2012; Pallas, Entwisle, Alexander, & Slutka, 1994). Indeed, the average academic skills of children in a class predict teachers' pace and level of instruction, as well as their academic expectations for students (Dreeben & Barr, 1988; Pallas et al., 1994), and increases in aggregate behavior problems tend to reduce instructional time (Carr et al., 1991). Thus, we may expect that, in pre-K classrooms with higher average incomes and lower proportions of children from low-income households, students may have more learning opportunities because teachers spend more time teaching, engage in more complex, student-centered instruction, and have more positive interactions with children.

Finally, the structure of pre-K classrooms may be impacted by classroom-economic composition. Increased classroom economic advantage predicts more time in free play, less time spent in routine activities like getting in line, cleaning up after meals, transitioning between activities, and more time spent in learning activities (Early et al., 2010). Thus, in less economically advantaged classrooms, children's academic skills may grow more slowly if they spend a smaller amount of time in learning activities and free play, both of which have been tied to early language, literacy, and math achievement (e.g., Cabell, DeCoster, LoCasale-Crouch, Hamre, & Pianta, 2013; Connor et al., 2005; Ginsburg, Lee, & Boyd, 2008).

It is critical to note that not all theories of peer effects posit benefits of classroom integration. Some models contend that being instructed with peers of similar SES, and presumably abilities, produces optimal achievement outcomes because teachers can direct their instruction, lesson plans, and materials to the predominant skill level (e.g., Hoxby & Weingarth, 2005). Similarly, theories of relative disadvantage, sometimes referred to as the "frog pond" perspective, highlight the negative consequences of socioeconomically integrated classrooms because children may be evaluated by their relative standing in the classroom, that is, being the "small frog in the pond" (Crosnoe, 2009; Marsh & Hau, 2003). It suggests that poor students may face greater competition for grades and more risks for stigmatization in economically integrated classes than in classrooms with similarly situated peers. Thus, a measure of the standard deviation or spread among students' economic backgrounds is an important compositional factor that may affect individual achievement.

Classroom-Economic Composition and Individual Achievement

Relations between peer-compositional characteristics and individual achievement were noted as early as 1966 in the Equality of Educational Opportunity Study, which is commonly referred to as "the Coleman report." Findings suggested that aggregate socioeco-

nomic characteristics at the school-level were more important in explaining individual student achievement than factors like school facilities, curriculum, or teaching quality (Coleman et al., 1966). Since then, an abundance of literature has documented associations between economic characteristics at the classroom- or school-level and individual students' achievement (see van Ewijk & Sleegers, 2010; Teddlie, Stringfield, & Reynolds, 2000 for reviews). A meta-analysis of 30 studies concluded that the average family income of peers has a moderate ($d \approx 0.25$ *SD*) positive association with individual achievement (van Ewijk & Sleegers, 2010).

Researchers attempting to account for omitted-variable bias (i.e., bias occurring when estimates do not control for important causal factors that are correlated with economic characteristics of the classroom and individual student achievement) obtained effect sizes that were smaller than studies making no such efforts (e.g., Hutchison, 2003; Rivkin, 2001; Strand, 1998). This is an important consideration given that children are not randomly assigned to classrooms or schools, and factors that lead children to a particular classroom or school may also be related to children's SES and development. Also, studies using different measures of classroom-economic composition often have conflicting results. This is especially evident when examining percent of the class/school that is low income as a measure of classroom-economic composition, which is most often operationalized as percent eligible for free- or reduced-price lunch (FRL). This measure is salient to the universal versus targeted pre-K debate because FRL eligibility is often used as a criterion for enrollment. Some studies have found negative associations between increased numbers of low-income peers and individual achievement (e.g., Cooley, 2010), but others uncover no links (e.g., Bankston & Caldas, 1998; Hanushek, Kain, Markman, & Rivkin, 2003). Greater inconsistencies and null effects in studies using school reports of FRL eligibility instead of direct reports of family resources are not surprising, given that it is an error-prone measure of economic disadvantage (Harwell & LeBeau, 2010). A goal of the current study was to examine links between classroom-economic composition and individual achievement using a measure derived from parental reports of household income.

It is noteworthy that studies linking classroom-economic composition to individual achievement have generally used school-aged samples. Some of the hypothesized mechanisms for peer effects, such as academic tracking, competition for grades, and self-evaluation based on one's standing relative to peers (e.g., Crosnoe, 2009), are less relevant in the preschool years. Accordingly, it is not clear that prior findings generalize to pre-K classrooms. The lone exception, a study using data from the National Center for Early Development and Learning (NCEDL; Reid & Ready, 2013), confounded income and maternal education, which limited its ability to inform pre-K policy because program eligibility is generally based on income, not education. For the present study, we controlled for other child, family, and classroom characteristics, including average maternal education level, to explore the unique role of classroom-economic composition in predicting achievement.

Nonlinearities in Relations Between Classroom-Economic Composition and Achievement

Existing research on classroom-economic composition and achievement is based on the assumption that the association is linear, which suggests that an increase in mean classroom income or the proportion of children from low-income households is related to the same achievement growth, regardless of the level of classroom economic characteristics. This presumes that moving from a classroom with a mean income level of $20,000 to a classroom with a mean income level of $30,000 predicts the same gains in achievement as moving from a classroom with mean income of $80,000 to one with $90,000. This assumption has not been empirically tested and is contrary to research on links between individual income and achievement that has found that increased income predicts larger achievement gains for children from low-income families (e.g., Duncan, Ziol-Guest, & Kalil, 2010).

Furthermore, evidence from the broader peer-effects literature suggests that nonlinear associations should be considered (e.g., Hoxby & Weingarth, 2005; Lazear, 2001; Neidell & Waldfogel, 2010). For example, studies have uncovered evidence of a "tipping point" of peer effects, whereby the positive effects of increasing numbers of more advantaged or skilled students are not evident until the proportion of skilled students in the class/school reaches a certain point (e.g., Lazear, 2001; Neidell & Waldfogel, 2010), though at least one study observes the opposite pattern—positive effects of high-achieving peers diminish as classes get more skilled (Zimmer & Toma, 2000). A novelty of the present study is its use of nonparametric modeling techniques to examine nonlinear links between classroom-economic composition and achievement.

Differences in Associations by Individual Ability

Research on economic composition and individual achievement has not included exploration of whether these relations depend on individual characteristics, despite evidence of such moderation in the broader peer-effects literature. For instance, several studies have shown that links between peers' academic abilities and achievement are strongest for the lowest performing students (Burke & Sass, 2013; Justice, Petscher, Schatschneider, & Mashburn, 2011; Zimmer & Toma, 2000). However, findings are mixed, with some studies failing to find moderation by initial skill level (Hanushek et al., 2003) and others showing harmful associations of lower skilled peers and achievement for students with more advanced academic skills (Imberman, Kugler, & Sacerdote, 2012, see also Mashburn, Justice, Downer, & Pianta, 2009). These studies were investigations of peer-academic ability as the compositional predictor. However, they guided our examination of moderation by individual ability in the present study of pre-K-classroom-economic composition's role in predicting achievement.

Research Aims

This study had three research aims, the first of which was to examine whether classroom-economic composition, as measured by average family income, standard deviation of classroom incomes, and percentage of students from low-income households, was associated with student achievement in pre-K. Particular attention was given to examining whether these links are nonlinear. The second aim was to estimate the size of associations between classroom-economic composition and academic skills across classrooms of varied economic composition. Third was to consider whether relations between classroom-economic composition and

achievement differed based on children's initial achievement. Based on prior research, we hypothesized that classroom advantage (i.e., higher aggregate-mean income and lower percentages of students from low-income households) would positively predict individual achievement, but we did not expect significant links between the standard deviation of classroom incomes and individual achievement. We also hypothesized threshold effects of classroom-economic composition, whereby classroom advantage would be more strongly associated with achievement in the most disadvantaged classrooms. Last, we predicted that the benefits of classroom advantage would be the most pronounced for children with less advanced academic skills at the start of pre-K.

Method

Participants

Data for this study were drawn from two prospective evaluations conducted by the NCEDL of state-funded pre-K programs: the Multi-State Study of Pre-Kindergarten (Multi-State) and the State-Wide Early Education Programs Study (SWEEP; Early et al., 2005). The Multi-State and the SWEEP included programs from 11 states[1] that had traditionally high rates of pre-K enrollment and were diverse in terms of geography, dominant type of pre-K model applied, and program characteristics. Sampled programs included a mix of universal and targeted pre-K, including Head Start (Dotterer, Burchinal, Bryant, Early, & Pianta, 2013). Six states contained only targeted programs (i.e., those using income requirements ranging from 110%–350% of the FPL), four had universal programs, and one state contained both. Across both the Multi-State and the SWEEP, 63% of children were enrolled in targeted programs and 37% in universal programs. Universal and targeted programs varied widely in terms of classroom-economic composition. Classroom aggregate income in universal programs averaged approximately $44,000, with a low of $8,000, a high of $85,000, and a standard deviation of about $20,000. On average, the percentage of pre-K students that were low-income in universal classrooms was just under half (48.4%), but ranged from 0% to 100%. Thus, the universal programs in this sample were economically heterogeneous. As expected, targeted programs were more disadvantaged and less economically integrated. Average aggregate income was about $25,000, though it ranged from $7,000 to $78,000,[2] and there was less variability in family incomes (e.g., standard deviation was less than $12,500). The classrooms in targeted programs were predominantly made up of students from low-income households (81.4%), with a range of 10%–100%.

The investigators of both studies randomly sampled sites (i.e., pre-K centers) within states, one classroom within each site, and four children within each classroom (see Early et al., 2005 for a description of the sampling process). Multi-State investigators used a multistage, random sampling process in 20 zip codes from each state, two sites from each selected zip code, one pre-K classroom from each selected site, and four pre-K children from each selected classroom. This random sample was stratified within-state by teacher education, program location (inside vs. outside of a school), and program type (full- vs. part-day programs). Pre-K data collection for the Multi-State took place during the 2001–2002 school year. Of the 335 sites contacted, 238 sites participated in the fall and two additional sites agreed to in the spring. SWEEP pre-K data collection occurred during the 2003–2004 school year. The investigators recruited a random sample of 100 pre-K sites per state, stratified by county or district. In total, 465 sites participated in the SWEEP in the fall, and 463 of those continued participation in the spring. In both studies, target students that left their pre-K classrooms between fall and spring assessments or dropped out of the study were replaced with randomly selected students from the same classroom. Four target children switched to another pre-K classroom within the same site. These children remained in the study, which added another four classrooms to the sample, and another four children in the original classrooms were added.[3] There were 142 students who dropped out of the study and were replaced with new students from the same classrooms. Both the Multi-State and the SWEEP employed the same measures and training, which allowed us to collapse the two datasets to address our research aims.[4] Thus, across the Multi-State and SWEEP, we analyzed data collected from 2,966 children, who ranged in age from 3.8 to 5.7 years of age ($M = 4.6$) at the start of pre-K, in 709 classrooms in 11 states for the current study.

Procedure

In the classrooms randomly selected for participation, families of all children received packets containing a consent form and a demographic questionnaire. Data from this demographic questionnaire were pooled across all children in the classroom to generate measures of classroom-economic composition. Four children were randomly selected as target children from all children in a classroom. Direct assessments of the target children's academic skills were performed. In addition, teachers completed questionnaires about the target children, and answered questions regarding their own educational background and classroom characteristics.

Direct assessments of target children's academic skills were conducted in the fall and spring of pre-K. Children who spoke a language other than English at home were given a portion of the *pre*LAS, an English language proficiency assessment, (Duncan & De Avila, 1998) to screen for English proficiency. Children who did not score at least 31 of 40 possible points and spoke Spanish at home ($n \approx 300$) were given assessments administered in Spanish (either using the Spanish-language equivalents of the English tests or by translating the tasks into Spanish). Children who did not pass the *pre*LAS and spoke a language other than Spanish at home were not assessed. All children who were administered direct assessments are included in our analysis. We combined the English and Spanish measures to retain the maximum sample of children

[1] The six states included in the Multi-State Study were California, Georgia, Illinois, Kentucky, New York, and Ohio. The five states included in SWEEP were New Jersey, Massachusetts, Texas, Wisconsin, and Washington.

[2] Although it is unusual that a "targeted" pre-K classroom would have an average income as high as $78,000, researchers have documented that there are often circumstances in which pre-K programs classified as targeted enroll students that are not low-income (Early et al., 2005).

[3] The classrooms that these students came from and entered into were not significantly different on the classroom measures collected in the fall and spring ($p = .1–.9$).

[4] Model noninvariance was tested by interacting an indicator for data set with our key variables of interest. There were no significant interactions, which further justified pooling the data.

and controlled for the language of assessment. This is consistent with similar preschool studies (e.g., Parrish & Howes, 2008; Wong, Cook, Barnett, & Jung, 2008), and the results of this study are robust to the exclusion of these children.

Measures

Literacy and language skills. Children's early literacy and language skills were assessed using three measures. The Multi-State and SWEEP (Early et al., 2005) assessed letter-identification skills by showing children a set of mixed capital and lowercase letters and asking them to identify as many letters as they could (Bryant, Barbarin, & Aytch, 2001). Scores on this assessment range from 0–26 (α = .97). The NCEDL Early Writing Task (NEWT) was used to assess children's emergent literacy skills (NCEDL, 2005). Children were asked to write their names, as name writing has been found to be an important indicator of early literacy and language knowledge (Bloodgood, 1999; Whitehurst & Lonigan, 1998). The proportion of the name the child could write legibly, ranging from 0–100%, was coded. Several research members coded 50 writing samples to establish reliability. The κ value had a mean range of .76–.89, and coders scored exactly the same or within one percentage point 87%–97% of the time. After establishing reliability, NEWT scores were coded by one researcher. Language skills for the English-speaking children were assessed using Dunn and Dunn's (1997) *Peabody Picture Vocabulary Test* (3rd ed.; PPVT-III; fall: α = .96, spring: α = .96) or the *Test de Vocabulario en Imagenes Peabody* (TVIP; fall: α = .92, spring: α = .93; Dunn, Padilla, Lugo, & Dunn, 1986) for the Spanish-speaking children. The raw PPVT-III and TVIP scores were used because the standard scores are standardized on different populations (English-speaking U.S. children vs. children in Puerto Rico and Mexico, respectively). These measures correlated at around .40. To create a combined measure of literacy and language, scores on these three assessments were standardized (within time) and averaged.[5] We used children's literacy and language scores from the spring assessment as the outcome. To reduce omitted-variable bias, fall scores were included as an independent variable to control for unmeasured, time-invariant differences in children and families that affect children's achievement and may be correlated with the key independent variables of interest (Chase-Lansdale et al., 2003).

Math skills. Investigators administered the Applied Problems subtest of the *Woodcock–Johnson III Tests of Achievement* (Woodcock, McGrew, & Mather, 2001; fall: α = .84, spring: α = .83) to measure children's math reasoning and problem-solving skills. It requires the child to analyze and solve math problems by performing relatively simple calculations. Children not passing the English language proficiency screener were given the *Batería Woodcock–Muñoz-Revisada: Pruebas de Aprovechamiento, Problemas Aplicados* (Woodcock & Sandoval, 1996; fall: α = .81, spring: α = .79). Raw scores on the Applied Problems assessment were used. In addition, children were shown teddy bears and asked to count them with one-to-one correspondence to measure emergent numeracy skills (Gelman & Gallistel, 1986). The highest number counted in the sequence was recorded, with a maximum score of 40. Scores on these two assessments were standardized (within time) and averaged to create a single measure of early math achievement. Correlations between the two measures averaged .45.

Classroom-economic composition. Measures of classroom-economic composition were derived from the demographic questionnaire that was administered to all children in target children's classrooms. Three measures of classroom-economic composition were considered. The first was a measure of the mean classroom family yearly income (scaled in $10,000 increments). All measures of income obtained from participants in the Multi-State, which were collected in 2001, were escalated to 2003 dollars to be consistent with the income measures from SWEEP, which were collected in 2003 (Early et al., 2005). The second measure reflects the percent of students in the classroom from low-income families (scaled in 10% increments), with low-income defined as having family income less than 200% of the FPL. We chose 200% as our low-income cutoff because states vary in income thresholds used to determine eligibility for pre-K subsidies, with cut-offs ranging from 100% to over 300% of the FPL. The majority of states use 185%–200% as the requirement (Barnett et al., 2014). Moreover, recent political oratory, including President Obama's plan for early education, has argued for the use of 200% as the income guideline for subsidized pre-K program enrollment (e.g., Office of the Press Secretary, 2013). We consider it important that, in addition to the 200% level, we tested 100%, 150%, and 185% of the FPL in our analyses and results were consistent across all specifications. Last, we tested variability in classroom income with a measure of the standard deviation of classroom incomes.

Child and family characteristics. Several child and family covariates were included in the models.[6] These were derived from interviews with the parents of the target children. We controlled for the target child's age, gender, time between assessments, whether the child was assessed in English or Spanish, and race/ethnicity, which was represented with dummy variables indicating whether the child was White (reference group), African American, Latino, or another race, which included Asian, Native American, and multiracial. Family structure was represented with an indicator of whether the target child lived with a single parent and a continuous measure of the number of children under 18 years of age living in the household. The highest level of maternal education was coded in three categories: (a) less than a high school degree (reference group), (b) a high school diploma or GED, and (c) a bachelor's degree or higher. Family income was measured continuously in 2003 dollars and is expressed in $10,000 increments. Last, 10 state indicators were included in the models to represent the state in which the target child resided to control for state effects.

Classroom characteristics. Several classroom characteristics were included in the models to reduce the likelihood that observed associations between classroom-economic composition and individual achievement were driven by other characteristics of pre-K classrooms found to be correlated with children's achievement and classroom SES. First, using data from the demographic questionnaire administered to all classroom students, a measure of aggregate classroom maternal education was created by averaging total

[5] Initially, individual assessments were modeled as separate outcomes. However, results were similar across assessments so we combined them into a single literacy and language measure. The same is true for math assessments.

[6] We tested for interactions between all covariates and our economic composition variables. None were significant.

years of maternal schooling across all children.[7] This was a central control variable in our efforts to identify the unique association between classroom-economic composition and achievement, as opposed to other aspects of classroom SES. Also using data from the demographic questionnaire, we accounted for the racial/ethnic composition of the classroom by including a measure of the percentage of students in the classroom who were African American, Latino, Asian, Native American, or multiracial. We also included dichotomous indicators for whether the program was full-day or part-day and whether the head teacher possessed at least a bachelor's degree. We also controlled for class size and teachers' years of teaching experience. We chose to include these structural aspects of classrooms because they have been identified as potential quality indicators that may relate to achievement, but also vary by students' SES (LoCasale-Crouch et al., 2007; Mashburn et al., 2008). The failure to control for these indicators could upwardly bias our estimates of peer effects. Other characteristics of classroom processes and interactions between teachers and students, such as instructional time and teachers' emotional supportiveness, were not included as controls because they are potential pathways through which classroom-economic composition may shape the development of academic skills (e.g., Creemers & Reezigt, 1996), and including these as controls could downwardly bias estimates of peer effects.

Data Analysis

To address our first research question, which considered whether classroom-economic composition is related to student achievement, we estimated nonparametric equations using general additive modeling (GAM; Hastie & Tibshirani, 1990) in SAS 9.4. GAM can be used to identify thresholds at which associations change by estimating the relation between a predictor and outcome without making assumptions about whether the nature of that relation is linear, quadratic, logarithmic, etc. Instead, functional form is determined empirically by the data. Specifically, GAM allowed the data to model the nonlinear relations between each of our three measures of classroom-economic composition and achievement after controlling for all covariates. GAM output provides a plot that gives accurate and reliable visual guidance as to the functional form that best characterizes associations and regions where thresholds exist (Setodji et al., 2012). GAM models did not account for the multilevel nature of the data. However, because clustering biases standard errors, not regression coefficients (Cohen, Cohen, West, & Aiken, 2003), and GAM was not being used to test the significance of parameter estimates, its use here was appropriate.

GAM requires researchers to make judgments about the location of thresholds and does not provide significance tests of parameter estimates. Thus, the validity of identified thresholds must be tested with other statistical methods. Accordingly, thresholds were tested using spline regressions, with each potential threshold constituting a spline knot. These parameterized models (one for mean income, one for standard deviation, and one for low-income percentage) allowed us to examine whether the magnitude of classroom-economic composition's associations with academic skills before and after the visually chosen thresholds significantly differed from each other. These models also controlled for all covariates.

Next, the statistical significance and size of relations between classroom-economic composition and achievement were estimated with hierarchical linear modeling. First, we estimated spring academic skills as a function of each of our economic composition measures separately, controlling for the child's fall scores, individual-level child and family characteristics, and other classroom characteristics. Children's fall scores were grand-mean-centered to reduce multicollinearity in the moderation models (Cohen et al., 2003). We included a classroom-level random effect to take into account the nesting of children within classrooms (Rabe-Hesketh & Skrondal, 2008).[8]

Our final research aim was to examine whether children's academic skills at baseline (the start of pre-K) moderated relations between classroom-economic composition and achievement. To answer this question, we added interactions between children's fall academic scores (grand-mean-centered) and our composition measures. The interactions were generated using the appropriate functional form identified in our initial GAM analyses.

All parameterized models, including those testing the validity of observed GAM thresholds, were estimated using "xtmixed" in Stata 13. Mixed-effects models take into account the nesting of children within pre-K classrooms. We constrained regression slopes to be equal across classrooms, but allowed intercepts to vary. By taking account of the clustering of children within classrooms, mixed-effects modeling provides more accurate parameter estimates and standard errors, thereby reducing the likelihood of Type-I errors (Bryk & Raudenbush, 1992). In addition to estimating models with classroom-economic composition variables separately, we also ran models with mean income and standard deviation of class incomes in the same model and standard deviation and low-income percentage in the same model to determine whether results changed when controlling for the variability in income. Results were robust to the inclusion of standard deviations. We could not estimate models with all three due to multicollinearity between mean income and percentage of students from low-income households. All model assumptions were tested using standard techniques (Cohen et al., 2003), and no violations were observed.

Missing Data

There were missing data in the combined data set, though all cases had valid data on some variables. The amount of missing data varied depending on whether the data came from the assessment of the target child, the interview with the target child's parents, the interview with the teachers, or the demographic questionnaire given to all families of children in the target children's classrooms. Missing data from the target children ranged from 7–9%, and variables created using data from their parents were missing in 0–16% of cases. There were very little data missing from teachers—only 2–6%.

[7] We could not create categorical measures of classroom-aggregate maternal education level because the individual parent responses from classroom peers were not available in the data sets, as explained in detail in the Missing Data section.

[8] We were unable to account for the nesting of children within zip code because zip code data were not available.

Missing data were not missing completely at random (MCAR) according to Little's test for MCAR. Having missing values was related to several other variables in our analyses, including achievement, classroom- and individual-level SES, and teacher characteristics. To retain participants without full data in our analyses, missing data were imputed in Stata 13 using the multiple imputation by chained equations technique to create 20 imputed datasets (Royston, 2004). All variables in our analyses, including those related to having missing values, were used in our imputation equations. Imputed data were used for all analyses reported in our results except GAM analyses, in which case listwise deletion was performed.

Every classroom in our sample had valid data on classroom-economic composition and other compositional characteristics drawn from the demographic questionnaire; no compositional characteristics were missing data variables. This is because the study investigators created these classroom compositional variables based on all available children in each classroom, no matter how many families responded. We were unable to impute data on the classroom compositional variables because the raw data from the demographic questionnaires were not provided in the Multi-State and SWEEP datasets (Early et al., 2005). The average response rates for the demographic questionnaires completed by the parents of all children in target classrooms averaged 70%, but ranged from 10–100%. It should be noted that the average classroom response rate was high compared with similar studies (e.g., Henry & Rickman, 2007; Justice et al., 2011; Mashburn et al., 2009).

Results

Descriptive statistics for the sample can be found in Table 1 and highlight the socioeconomic diversity of the sample, both at the child and classroom level.

Nonlinearities in Associations Between Classroom-Economic Composition and Achievement

GAM diagnostics suggested that the functional form of relations between achievement and the standard deviation of classroom income were linear. There were nonlinearities in associations between academic achievement and the other two economic composition measures (i.e., mean income and low-income percent). With respect to literacy and language skills, increases in the percentage of students from low-income households had a weak association with skills until 25% of the class were low income, at which point a strong negative relation emerged. The association plateaued at 45% low-income. The relation between literacy and

Table 1
Selected Descriptive Statistics for Sample (N = 2,966)

Variable	*M* or %	*SD*
Academic outcomes		
Literacy and language skills		
Spring	−.01	.77
Fall	−.02	.81
Math skills		
Spring	−.02	.87
Fall	−.03	.88
Child-level characteristics		
Age in years at spring assessment	5.05	.32
Child is male	49.19%	.50
Assessed in spring in Spanish	11.63%	.32
Child race		
White	40.99%	.49
African-American	18.48%	.39
Latino	26.67%	.44
Other/multiracial	13.86%	.35
Family yearly income	$32,939.58	$25,529.09
Child lives in single parent home	40.44%	.49
Maternal ed.		
Less than high school	18.80%	0.39
High school	63.58%	.48
Bachelor's or greater	17.62%	.38
Number of siblings	2.43	2.55
Classroom-level characteristics		
Mean class income	$32,285.19	$18,024.89
% Of class that is low-income (<200% FPL)	69.18%	.31
Mean years of maternal education	12.77	1.36
Full-day prekindergarten	43.21%	.49
Teacher has bachelor's degree	70.75%	.45
Teacher's years of experience	13.22	9.23
Class size	18.52	5.56
% Of class racial/ethnic minority	59.31%	.37

Note. Descriptive statistics were calculated using imputed data. Academic outcomes are in standard score units. Classroom-level characteristics were calculated at the individual level for 2,966 children in 709 classrooms.

language skills and mean income was linear. On the other hand, GAM diagnostics revealed nonlinearities in links between math achievement and both mean classroom income and percent low income. Specifically, mean income was not related to math scores until mean levels reached about $22,500, at which point it had positive links until approximately $62,500, when the association plateaued. When looking at links between low-income percent and math, children's math scores remained relatively flat as the percentage of students from low-income households in classrooms increased, until the classroom reached just over 50% low income, at which point the relation grew much steeper in the negative direction. It again weakened after more than 70% of the class was low income. These nonlinearities were best fit by spline functions. The results of the parameterized models showed that observed thresholds were valid, meaning the size of relations between achievement and mean income and percent low income differed significantly before and after the thresholds (25% and 45% for low-income percent and literacy and language, $22,500 and $62,500 for mean income and math, and 52.5% and 72.5% for percent low income and math).

Associations Between Classroom-Economic Composition and Individual Achievement

Associations between the standard deviation of classroom income and individual achievement were never significant, so for parsimony's sake these were dropped from the model and are not discussed further. Table 2 presents unstandardized regression coefficients from models assessing associations between academic skills and average family income in the classroom, controlling for fall scores as well as a host of other child, family, and classroom characteristics. In the text below, we report standardized effect sizes to provide more meaningful interpretation of results. As seen in Table 2, when predicting literacy and language skills, mean-classroom income is represented with a linear measure. Because GAM analyses revealed nonlinear links between mean income and

Table 2
Mixed-Effects Regression Predicting Spring Academic Skills With Mean Classroom Income

	Literacy and language		Math	
Variable	Coefficient	(*SE*)	Coefficient	(*SE*)
Linear				
Mean class income	.004	(.01)		
Spline				
Mean class income <$22,500			−.05	(.04)
Mean class income = $22,500–$62,500			.07***	(.02)
Mean class income >$62,500			−.04	(.04)
Child characteristics				
Fall skills	.71***	(.01)	.67***	(.01)
Age	.10**	(.03)	.24***	(.04)
Days between assessments	.002***	(.00)	.002***	(.00)
Child is male	−.03	(.02)	−.06**	(.02)
Race				
African American	−.03	(.03)	−.04	(.04)
Latino	.01	(.04)	.02	(.04)
Other race	−.01	(.03)	.02	(.04)
Assessed in Spanish	−.32***	(.03)	−.07	(.05)
Family income	.01	(.01)	.01	(.01)
Single-parent home	−.03	(.02)	−.04	(.03)
Maternal education				
High school	.04	(.02)	.01	(.03)
Bachelor's degree	.05	(.04)	.04	(.05)
Number of siblings	−.02**	(.01)	.01	(.01)
Classroom characteristics				
Full-day prekindergarten	.02	(.03)	.02	(.03)
Teacher has bachelor's degree	.06*	(.03)	.05	(.03)
Teacher experience	−.001	(.001)	−.001	(.001)
Class size	.002	(.002)	.003	(.002)
% Minority	.001	(.05)	.01	(.06)
Mean classroom maternal education	.004	(.01)	.01	(.02)
Intercept	−.94***	(.21)	−1.70***	(.28)
Random-effects parameters				
Within-classroom variance	.29***	(.0001)	.15***	(.00004)
Between-classroom variance	.01***	(.001)	.02***	(.0002)

Note. $N = 2{,}966$. There were 1,174 children in 276 classrooms with mean class incomes of <$22,500, 1,505 children in 362 classrooms with mean class incomes of $22,500–$62,500, and 287 children in 71 classrooms with mean class incomes of >$62,500. Coefficients are unstandardized. Dummy variables indicating children's state of residence were also included in models, but results are not shown here. Categories for child's race are compared with the omitted White group. Maternal education categories are compared with the omitted "below high school" group.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

math scores, mean income is represented with a spline specification with two thresholds: one at $22,500 and one at $62,500. Thus, there are three mean-income terms. The first term represents the slope of the relation between mean class income and math achievement below $22,500. The second term is the slope of the relation between mean income and math between $22,500 and $62,500, and the last term is the slope of the relation after the $62,500 threshold is reached. Table 3 shows results of parameterized models looking at the percent of children in the classroom who come from low-income households as our indicator of classroom-economic composition. Spring achievement was predicted using spline terms with thresholds at 25% low income and 45% low income for literacy and language and at 52.5% and 72.5% for math. Similar to mean income, the three coefficients represent, respectively, the slope before the first threshold, between the two thresholds, and after the second threshold.

Results indicate that average classroom income was unrelated to children's literacy and language skills (see Table 2). However, there was a negative relation between the percentage of preschoolers who come from low-income households in a classroom and literacy and language skills (see Table 3), but only within the 25%–45% low-income range. More specifically, once a quarter of the preschoolers in the class came from low income, 10% increases in the percentage of students from low income were related to a .06 *SD* decrease in literacy and language achievement. Beyond the low-income threshold of 45% of the class, further increases were unrelated to literacy and language achievement. While we do not discuss relations between the covariates and achievement for parsimony's sake, it is important to note that although there were some significant associations between the covariates and achievement (most notably the strong association between fall and spring scores), many child-level characteristics, including children's own household income, and all of the other classroom-level

Table 3
Mixed-Effects Regression Predicting Spring Academic Skills With Low Income Percent of Class

	Literacy and language		Math	
Variable	Coefficient	(*SE*)	Coefficient	(*SE*)
Spline				
<25% class low income	.03	(.02)		
25–45% class low income	−.05*	(.02)		
≥45% class low income	.01	(.01)		
<52.50% class low income			−.01	(.01)
52.50–72.50% class low income			−.08**	(.02)
≥72.50% class low income			.03	(.02)
Child characteristics				
Fall skills	.71***	(.01)	.67***	(.01)
Age	.10**	(.02)	.24***	(.04)
Days between assessments	.002***	(.00)	.002***	(.00)
Child is male	−.03	(.02)	−.06**	(.02)
Race				
African American	−.03	(.03)	−.04	(.04)
Latino	.01	(.04)	.01	(.04)
Other race	−.004	(.03)	.02	(.04)
Assessed in Spanish	−.32***	(.03)	−.07	(.04)
Family income	.01	(.01)	.01	(.01)
Single-parent home	−.03	(.02)	−.03	(.03)
Maternal education				
High school	.04	(.02)	.01	(.03)
Bachelor's degree	.04	(.04)	.03	(.04)
Number of siblings	−.02**	(.01)	.01	(.01)
Classroom characteristics				
Full-day prekindergarten	.01	(.03)	.02	(.03)
Teacher has bachelor's degree	.06*	(.03)	.05	(.03)
Teacher experience	−.001	(.001)	−.0003	(.001)
Class size	.002	(.002)	.003	(.002)
% Minority	−.001	(.05)	−.001	(.06)
Mean classroom maternal education	.01	(.01)	.01	(.01)
Intercept	−.99***	(.24)	−1.73**	(.30)
Random-effects parameters				
Within-classroom variance	.15***	(.00004)	.29***	(.0001)
Between-classroom variance	.02***	(.0002)	.01***	(.001)

Note. $N = 2,966$. There were 388 children in 95 classrooms with < 25% low-income students, 325 children in 80 classrooms with 25%–45% low-income students, and 2,253 children in 534 classrooms with mean class incomes of ≥45% low-income students. There were 856 children in 209 classrooms with <52.5% low-income students, 397 children in 94 classrooms with 52.5%–72.5% low-income students, and 1,713 children in 406 classrooms with mean class incomes of ≥45% low-income students. Coefficients are unstandardized. Dummy variables indicating children's state of residence were also included in models, but results not shown here. Categories for child's race are compared to the omitted group of white. Maternal education categories are compared to the omitted group of below high school.
* $p < .05$. ** $p < .01$. *** $p < .001$.

characteristics, including mean maternal education, did not predict literacy and language scores above and beyond other predictors in the model. This was true for models predicting math skills as well. This is likely because our models controlled for achievement skills in the fall of pre-K.

Moving on to math, increases in mean classroom income positively predicted math skills, but only when the increases occurred between \$22,500 and \$62,500 (see Table 2). In that range, \$10,000 increases of aggregate classroom income were related to increases of .08 *SD* in math skills. Under \$22,500 and over \$62,500, income increases did not relate to improvements in math skills. There were also associations between low-income percent and math skills (see Table 3). As results show, increases in the percentage of children from low-income households have negative links with math achievement when preschool classrooms have between 52.5% and 72.5% of children from low-income households. Put differently, increases in the percentage of children from low-income households did not predict declines in math skills until about half of the classroom came from low-income homes. At that point, further increases in the proportion of students from low-income households related to lower math skills (.09 *SD* decrease in math achievement for every 10% increase in students from low-income households). This association plateaued when roughly 72.5% of the students had low income, at which point increasing proportions of students from low-income households were not related to math skills declines.

Moderation by Children's Initial Academic Ability

Tables 4 and 5 present the results of models examining whether associations between classroom-economic composition and academic skills were moderated by children's academic skills at the start of pre-K. All interaction models controlled for all covariates (listed in the notes of the tables), but for parsimony's sake only the main effects and interaction terms are presented in the tables. Interactions were tested using the GAM-identified functional forms.

Results revealed few significant interactions between individual skills and classroom-economic composition, with one notable exception. For early literacy and language skills, relations between mean classroom income and academic skills were weaker for children who began pre-K with better skills (see Table 4). Figure 1 illustrates the associations between mean classroom income and literacy and language skills for children with high (scoring 1 *SD* above the mean), average (the mean) and low (1 *SD* below the mean) fall skills. As initial literacy and language skills increase, the relation between mean class income and skills diminished. For instance, while the main-effect model (see Table 2) showed no link between aggregate income and literacy and language skills, moderation results revealed that increases in aggregate classroom income predicted significant growth in literacy and language skills for students whose fall skills were 0.7 *SD* below the mean or less. There were no significant relations between aggregate classroom income and literacy and language skills for children with average or high fall skills. There was also evidence of initial skill moderation of the link between percent low income and literacy and language skills, though only for income changes greater than 45% (see Table 4). However, the practical significance of this interaction is questionable because simple slope tests revealed that associations between percent low income and literacy and language were not significant, even for children falling as much as 1 *SD* above or below the mean.

Table 4
Moderation of Association Between Mean Classroom Income and Spring Academic Skill Level by Children's Fall Skill Level

Variable	Literacy and language: Coefficient	Literacy and language: (*SE*)	Math: Coefficient	Math: (*SE*)
Linear				
Main effects				
Mean class income	.01	(.01)		
Fall skills	.80***	(.03)		
Interactive effect				
Mean class income × skills	−.03***	(.01)		
Spline				
Main effects				
Mean class income < \$22,500			−.05	(.04)
Mean class income = \$22,500–\$62,500			.07***	(.02)
Mean class income > \$62,500			−.05	(.05)
Fall skills			.58***	(.08)
Interactive effects				
Mean Class Income < \$22,500 × Skills			.05	(.04)
Mean Class Income = \$22,500–\$62,500 × Skills			−.02	(.01)
Mean Class Income > \$62,500 × Skills			.03	(.04)
Intercept	−1.01	(.21)	−1.70***	(.28)
Random-effects parameters				
Within-classroom variance	.15***	(.00004)	.29***	(.0001)
Between-classroom variance	.02***	(.0001)	.01***	(.001)

Note. $N = 2{,}966$. Coefficients are unstandardized. Models included all covariates listed in Table 2 and the state indicators.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Table 5
Moderation of Association Between % of Class Low Income and Spring Academic Skills by Children's Fall Skill Level

Variable	Literacy and language		Math	
	Coefficient	(*SE*)	Coefficient	(*SE*)
Spline				
Main effects				
<25% class low income	.02	(.03)		
25–45% class low income	−.07**	(.03)		
≥45% class low income	.01	(.01)		
<52.50% class low income			−.02	(.01)
52.50–72.50% class low income			−.07**	(.03)
≥72.50% class low income			.03	(.02)
Fall skills	.61***	(.05)	.61***	(.04)
Interactive effects				
<25% Class Low Income × Skills	.001	(.03)		
25–45% Class Low Income × Skills	.03	(.03)		
≥45% Class Low Income × Skills	.02*	(.01)		
<52.50% Class Low Income × Skills			.01	(.01)
52.50–72.50% Class Low Income × Skills			.03	(.03)
≥72.50% Class Low Income × Skills			−.03	(.02)
Intercept	−.95***	(.24)	−1.70***	(.30)
Random-effects parameters				
Within-classroom variance	.15***	(.00004)	.29***	(.0001)
Between-classroom variance	.02***	(.0001)	.01***	(.001)

Note. $N = 2{,}966$. Coefficients are unstandardized. Models included all covariates listed in Table 2 and the state indicators.
$^{*} p < .05$. $^{**} p < .01$. $^{***} p < .001$.

Discussion

Mounting evidence on school success demonstrates that children who begin school with strong early literacy, language, and numeracy skills are more likely to experience continued academic success compared with less prepared peers (e.g., Duncan et al., 2007). Thus, it is vital to understand predictors of early academic skills and to pinpoint high-risk groups and potential levers for prevention and intervention. Results from this study suggest that even after controlling for children's achievement at the start of the pre-K year, increased mean income and smaller percentages of

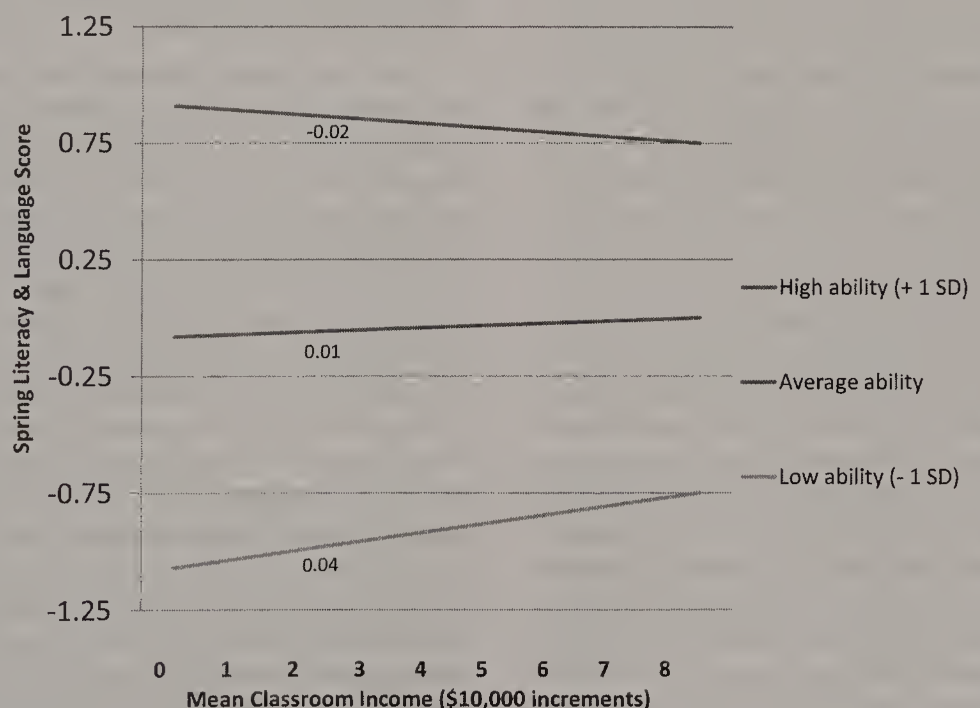


Figure 1. Interaction between mean classroom income and fall language and literacy skills. This figure presents slope coefficients (unstandardized) on mean classroom income for children falling 1 *SD* above, at, and 1 *SD* below the mean of fall literacy and language skills. See the online article for the color version of this figure.

students from low-income households in pre-K classrooms have positive relations with preschool children's achievement, particularly math achievement.

Our findings, which are consistent with earlier scholarship on children in primary and secondary school (see van Ewijk & Sleegers, 2010), extend this literature by focusing on preschool. Results also illustrate that frog-pond theories of peer effects are not relevant to preschool children. According to frog-pond theories, increases in classroom income and decreases in disadvantaged students would be harmful to the achievement of preschoolers from low-income households because more economically advantaged peers would raise competition for positive evaluations and cause the disadvantaged students to feel stigmatized and inadequate. Studies finding increased economic integration to be harmful to the achievement of children from low-income households have, notably, focused on high school students (e.g., Crosnoe, 2009). The processes purportedly driving those findings (e.g., competition for grades, stigmatization, and negative self-appraisals) seem less relevant in early childhood because grades are often not assigned, children tend to be unaware of their relative socioeconomic standing, and they are less likely to make negative self-evaluations in reference to peers (Suls, 1986).

Second, findings also suggest that links between both mean classroom income and percentage of students from low-income households and achievement are nonlinear. The positive association between aggregate classroom income and math scores starts at just above $20,000 aggregate income and tapers at around $60,000. Similarly, increases in the percentage of children from low-income households in pre-K classrooms relate to decreased math scores when the class is between approximately 50%–70% low income and decreased literacy and language scores when the class is between 25%–45% low income. These findings suggest that there is a "range of action" within which modifications in economic composition will produce changes in achievement, whereas increases or decreases in economic advantage under or over those thresholds have no links to achievement. Similarly, changes in the proportion of children from low-income households have no relation to achievement in either pre-K classrooms with high concentrations of disadvantaged students or with very few students from low-income households. Results add to the limited literature exploring threshold effects or nonlinearities when examining peer effects on achievement (e.g., Burke & Sass, 2013; Hoxby & Weingarth, 2005).

Third, we found little evidence that children's academic skills at the start of the pre-K year moderate relations between classroom-economic composition and improvements in academic skills over the preschool year. With the exception of mean classroom income and literacy and language skills, increasing classroom economic advantage within the "range of action" has similar benefits to achievement for all children, regardless of their skills at the start of the school year. To the authors' knowledge, this is one of the few studies to explore the moderating effects of individual ability on associations between achievement and classroom-economic composition, as opposed to peers' aggregate ability.

The results obtained from this study are important additions to the literature on peer-compositional effects because children's economic circumstances are a primary criterion used to determine eligibility for targeted pre-K programs. Hence, establishing links between classroom-economic composition and children's development in preschool is highly relevant to current policy considerations regarding whether public funds are better invested in targeted or universal pre-K programs. These results suggest that universal pre-K may narrow economic disparities in early achievement better than targeted programs, though universal pre-K access does not guarantee economically integrated classrooms because things like residential segregation and parental preferences for private preschool could lead to economically homogenous universal classrooms.

It is important to note that the effect sizes between both measures of classroom-economic composition and achievement during the pre-K year are small, 0.08 *SD* per $10,000 increase in income between $22,500 and $62,500 for math and 0.06 and 0.09 *SD* per 10% decrease in students from low-income households for literacy and language (between 25% and 45%) and math (between 52.5% and 72.5%), respectively. These effect sizes, however, are not inconsistent with prior studies of classroom SES and achievement that address omitted-variable bias (e.g., Hutchison, 2003; Rivkin, 2001). To put these results in perspective using the average targeted and universal classrooms contained in the Multi-State/SWEEP (Early et al., 2005), consider two otherwise similarly situated children with average achievement, except one is in a targeted classroom (average family income is $25,000 and 81% of the children are low-income) and the other in a universal classroom ($44,000 average income and 48% of children are low-income). We would expect that the child in the universal classroom would see math-skills growth over the pre-K year almost 0.2 *SD* greater than his or her peer in the targeted pre-K. We would not expect to see significant differences in literacy and language scores, because the difference in low-income rate between the two hypothetical classes is not within the 25%–45% range of action for associations between percent low income and literacy and language.

Differences in Results for Literacy and Language Versus Math Outcomes

There were some differences between the results for language and literacy and math. First, links between economic composition and skills were stronger for math than they were for literacy and language. Indeed, there was no association between aggregate income and literacy and language skills in main-effects models, though moderation models revealed that mean income was related to literacy and language skills for lower-achieving students. Next, threshold effects of classroom-economic composition were different for literacy and language versus math skills; there were none for links between mean income and literacy and language, and the location of thresholds occurred earlier for low-income percent and literacy than for math. Further, findings that children's initial ability levels moderated associations between classroom-economic composition and achievement applied only to early literacy and language skills.

These results reinforce developmental theory and research highlighting the differences in early language, literacy, and math skills development. For example, math skill acquisition requires active learning environments and direct, intentional instruction and is not as conducive to learning via modeling as is language/literacy development (Ginsburg et al., 2008). Preschool curriculum tends to be dominated by literacy instruction. Indeed, research has

shown that preschool teachers are more disinclined to teach math concepts than literacy/language skills (Blevins-Knabe et al., 2000; Ginsburg et al., 2008), and this may be especially so in classes with increased behavioral and learning problems. In addition, reliance on teacher-centered instruction tends to be greater in classrooms with high levels of disadvantage (Stipek, 2004), which may negatively impact math learning more than literacy (Ginsburg et al., 2008). We may have observed stronger links between classroom composition and math skills because, as classrooms become more highly disadvantaged, math instruction may be the first content area that preschool teachers eliminate. On the other hand, we may not have observed relations between classroom-economic composition and literacy and language because instruction in these areas is such a core, basic part of preschool curricula that classroom-economic composition may have smaller effects on the amount and/or quality of instruction that students receive. Moreover, early literacy and language development may be less negatively impacted by the teacher-centered instruction more common in classrooms with greater numbers of disadvantaged children (e.g., Ginsburg et al., 2008; Stipek, 2004), which might reduce classroom-economic compositional effects on early literacy and language skills.

Second, home environments may exert greater influence over literacy and language than math skills (Blevins-Knabe et al., 2000; LeFevre et al., 2009). Parents are more likely to engage in formal and informal literacy activities with young children through conversation, shared book reading, singing the alphabet song and other rhymes, and teaching letters than they are numeracy activities (Blevins-Knabe et al., 2000; Tudge & Doucet, 2004). There tend to be fewer math learning opportunities in children's home environments. Math skill acquisition may be more strongly influenced by classroom composition because children rely more on formal instruction in these settings. Thus, diminished quantity/quality of math instruction due to increased classroom disadvantage would disrupt math learning to a greater extent than literacy or language development, which is a strong focus in the home environment as well.

Similar processes may be driving the moderation by baseline skills of links between economic composition and literacy and language skills. Children with high initial literacy and language skills are likely receiving more frequent, rich, and complex language and literacy interactions at home (e.g., Bracken & Fischel, 2008; Brandt, 2001; Weigel, Martin, & Bennett, 2006). Children with average and below average literacy and language skills, on the other hand, may rely more on pre-K to foster growth in this domain, which could explain why we uncover stronger links between pre-K classroom-economic composition and literacy and language skills for children with less advanced skills. Math, on the other hand, may be primarily learned at school regardless of individual differences in math skills because children have more limited opportunities for math learning at home (Tudge & Doucet, 2004). Further, because preschool teachers are generally less apt to engage in math instruction (Ginsburg et al., 2008), math skills may be more highly compromised by increased behavioral or learning problems in more disadvantaged classrooms. In contrast, even in classrooms characterized by increased academic and behavior problems, teachers may still provide enough literacy instruction to promote literacy development for those who already have mastered basic skills, whereas the children who need more intensive teaching (the lower achievers) are disproportionately harmed by the limited opportunities for structured learning and peer modeling. Alternatively, there may have been ceiling effects on the literacy and language assessments for the children who entered pre-K with high literacy and language skills. Thus, diminished associations between classroom-economic composition and literacy and language skills as initial skill level increases may be attributable to limited literacy and language gains for these students due to their mastery of letter identification, rhyming, and name writing at the start of pre-K.

Threshold Effects of Classroom-Economic Composition on Achievement

The threshold effects of peer-economic composition observed in this study make a novel contribution to the literature on peer effects because prior studies have focused primarily on peer ability. The two thresholds, which defined the range of action, differ from prior studies exploring nonlinearities in peer effects. GAM allowed us to identify these thresholds because it is nonparametric and flexible and provided important guidance on where the thresholds occurred. Previous studies have used more rigid approaches, such as imposing a quadratic function (e.g., Zimmer & Toma, 2000) or using a priori thresholds (e.g., Neidell & Waldfogel, 2010), which may obscure important thresholds that do not fit the specific nonlinear function or thresholds that have been identified a priori.

Next, in cases in which nonlinearities were identified, both a tipping point and a point of fade-out were observed. Specifically, negative links between low-income percent and literacy and language and math scores were not observed until at least 25% and 52.5% of the class were students from low-income households, respectively. The positive relation between average classroom income and math achievement kicked in only at the $22,500 threshold. These tipping-point or critical-mass effects of classroom-concentrated disadvantage or economic integration have long been theorized (e.g., Crane, 1991; Johnson, Ladd, & Ludwig, 2002), and this investigation supports their existence with respect to peer-economic effects on early achievement. In addition to tipping points, we saw links between classroom-economic composition and achievement fade out after a certain level of mean income/economic integration was achieved, thereby providing important information about the levels of economic integration that may be beneficial for children's academic skills.

Threshold effects may represent the points after which teachers are no longer able to adequately address children's academic or behavioral issues without affecting the learning in the classroom. For example, teachers can address occasional learning or behavioral problems without having to sacrifice the amount, quality, or content of instruction or educationally enriching interactions. However, when the percentage of students from low-income households reaches the critical mass, further increases in the percentage of disadvantaged students may alter classroom climate. Fade-out of links between percentage of students from low-income households and achievement may occur when classroom problems become so pervasive that further increases in the prevalence of at-risk peers have little practical impact on teachers and students, and hence learning, in the class. It is interesting to note that the critical mass before which negative associations between percent low income and achievement become apparent differs for literacy

and language and math. Neidell and Waldfogel (2010) found a similar pattern of differences in the location of threshold effects of classroom composition on reading and math achievement.

In this study, the percent-low-income compositional variable is a stronger predictor of achievement than is mean classroom income. Low-income percent reflects the concentration of economically disadvantaged students in the class, which may be important because poverty places children at especially increased risk for academic and behavioral difficulties (Magnuson & Votruba-Drzal, 2009). Aggregate classroom income, on the other hand, reveals nothing about the ratio of disadvantaged to more advantaged children. For example, two classrooms with mean incomes of $33,000 may have entirely different climates. One may contain three students living in poverty with family incomes of $5,000 annually and seven middle-income students in families earning $45,000. The second may contain eight poor students with family incomes of $20,000 and two upper-income students whose families make $85,000. This classroom has more concentrated disadvantage than the first. We might expect that adding or subtracting one additional disadvantaged child to the latter classroom may have little impact on the climate if behavioral and learning problems are already so intense that the teacher and children are overwhelmed. But in the first classroom, changes in aggregate income due to adding or subtracting a highly disadvantaged child may result in significant changes in achievement by appreciably altering the learning climate. The different nature of these compositional variables should guide future research on peer effects. In particular, measures capturing the proportion of poor versus nonpoor students may be the best indicators of economic composition. But if it is expected that income effects exist more broadly across the income distribution, using aggregate income may be more appropriate.

Situating Results in Broader Research on Pre-K

This study may also provide insight for understanding variation in pre-K-program evaluations. For instance, universal pre-K studies, such as Gormley, Phillips, and Gayer (2008) and Gormley and colleagues' (2005) studies of Oklahoma's universal pre-K program and Weiland and Yoshikawa's (2013) study of universal pre-K in Boston, have found moderate to large (.40—1 *SD*) benefits (see also Henry et al., 2003). These effects are substantially larger than those uncovered in nationally representative studies of center-based preschool, which include private center-based preschools in addition to universal and targeted programs, and studies of targeted pre-K, including Head Start (range = 0.10–0.25 *SD*; e.g., Magnuson & Waldfogel, 2005; Votruba-Drzal et al., 2013; Zill et al., 2003). Our findings suggest that increased classroom-economic advantage in universal pre-K programs may play some role in the relatively larger effect sizes obtained from universal pre-K evaluations. Though it is important to note that there are other differences between the universal and targeted programs that may be important as well, including curriculum, program philosophy, teachers' qualifications and salary, and evaluation design. Of course, universal pre-K is not synonymous with economically integrated classrooms; children can be provided universal pre-K in highly segregated classrooms. However, the data from our study and the Boston pre-K program (Weiland, 2013) show that universal programs do result in, at least some, socioeconomically diverse classrooms.

Next, these results suggest that universal or socioeconomically integrated pre-K programs may be more beneficial for the early academic development of economically disadvantaged preschoolers than targeted preschool. Of course, we must acknowledge the potential negative implications that increased integration may have on middle- and upper-income children who are enrolled in preschools with classmates of similar family incomes. All else equal, our findings indicate that enrolling these children in more economically diverse classrooms, if the counterfactual is enrollment in a preschool with predominately advantaged students, may result in fewer gains in their academic skills during the pre-K year. Our results suggest, which is important to note, that as long as preschool classrooms do not exceed 25% low income and mean income does not drop below $62,500, adding disadvantaged students to upper-income classrooms will not negatively impact learning.

Limitations

Limitations of the current study must be acknowledged. First, the data are correlational. We made significant efforts to control for endogeneity bias by including a lagged measure of children's achievement and several child, family, and classroom covariates. Yet results cannot be interpreted as causal. Second, only four children per classroom were included in the study. Third, response rates for the aggregate classroom-composition variables averaged 70%. Thus, although the 70% response rates for compositional variables are better than those observed in many peer-effects studies, almost one third of data at the classroom-level are missing, which may introduce substantial measurement error into our compositional variables. However, the Multi-State and SWEEP (Early et al., 2005) are two of the few studies of preschool children that contain measures of classroom-compositional variables. Accordingly, the use of these measures marks an important contribution to the literature. Fourth, these data are now almost two decades old and are not nationally representative, which limits the generalizability of findings.

Fifth, GAM analytic techniques are heavily data driven, thus it is important for future studies to validate the threshold effects uncovered in this study. This is the only way to verify that our results capture true differences in associations between classroom-economic composition and early math achievement and are not due to idiosyncrasies in the sample used in this study.

Sixth, we have been unable to disentangle whether our results are driven by classroom characteristics or neighborhood characteristics, because they are probably related. Our data did not include information on children's neighborhoods, so it is impossible to explore how many of the observed relations were attributable to neighborhood-socioeconomic integration as opposed to integration within the classroom. The confound between economic characteristics in neighborhood and classroom, however, is likely less problematic in preschool than in elementary and high school because children are not districted to preschools in their neighborhood, and families commonly cross school-district lines and travel great distances for early education programs (see, e.g., Gordon & Chase-Lansdale, 2001 finding the geographic area most accurately capturing the market for center-based care to be 25 miles from

children's homes). Finally, data limitations prevented us from considering the mechanisms that mediate links between classroom-economic composition and individual achievement, such as aggregate academic and behavioral skills, teaching practices, and peer modeling. Future researchers should empirically test pathways through which classroom-economic composition relates to achievement in preschool.

Conclusion

In conclusion, the results from this study provide new evidence of relations between classroom-economic integration and children's academic skills in pre-K. These findings extend the compositional peer-effects literature by showing that the positive links between increased economic advantage at the classroom level and individual achievement that have been documented with older children exist in preschool populations as well. Furthermore, the results may have implications for the effectiveness of targeted versus universal pre-K programs in promoting children's school readiness. Universal programs enrolling middle- and upper-income children in addition to economically disadvantaged children may be more effective than targeted pre-K in promoting the early achievement of preschoolers from low-income households. Additional research is necessary, however, to identify causal effects and to pinpoint the mechanisms by which increased classroom advantage predicts enhanced academic achievement. A thorough understanding of when, why, and for whom economic integration in classrooms and schools is beneficial is necessary to inform program and policy decisions regarding the types of early education programs that will best prepare all children to succeed in kindergarten and beyond.

References

Arnold, D. H. (1997). Co-occurrence of externalizing behavior problems and emergent academic difficulties in young high-risk boys: A preliminary evaluation of patterns and mechanisms. *Journal of Applied Developmental Psychology, 18,* 317–330. http://dx.doi.org/10.1016/S0193-3973(97)80003-2

Bankston, C. L., III, & Caldas, S. J. (1998). Family structure, schoolmates, and racial inequalities in school achievement. *Journal of Marriage and the Family, 60,* 715–723.

Barnett, W. S., Carolan, M. E., Squires, J. H., Brown, K. C., & Horowitz, M. (2014). *The State of Preschool 2014: State Preschool Yearbook.* Rutgers, NJ: National Institute for Early Education Research. Retrieved from http://nieer.org/sites/nieer/files/yearbook2014_full.pdf

Battistich, V., Solomon, D., Kim, D. I., Watson, M., & Schaps, E. (1995). Schools as communities, poverty levels of student populations, and students' attitudes, motives, and performance: A multilevel analysis. *American Educational Research Journal, 32,* 627–658.

Blevins-Knabe, B., Austin, A. B., Musun, L., Eddy, A., & Jones, R. M. (2000). Family home care providers' and parents' beliefs and practices concerning mathematics with young children. *Early Child Development and Care, 165,* 41–58. http://dx.doi.org/10.1080/0300443001650104

Bloodgood, J. W. (1999). What's in a name? Children's name-writing and literacy acquisition. *Reading Research Quarterly, 34,* 342–367. http://dx.doi.org/10.1598/RRQ.34.3.5

Bracken, S. S., & Fischel, J. E. (2008). Family reading behavior and early literacy skills in preschool children from low-income backgrounds. *Early Education and Development, 19,* 45–67. http://dx.doi.org/10.1080/10409280701838835

Brandt, D. (2001). *Literacy in American lives.* New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511810237

Bryant, D., Barbarin, O., & Aytch, L. S. (2001). *Naming letters.* Unpublished measure, University of North Carolina at Chapel Hill, Frank Porter Graham Child Development Center.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Bub, K. L., McCartney, K., & Willett, J. B. (2007). Behavior problem trajectories and first-grade cognitive ability and achievement skills: A latent growth curve analysis. *Journal of Educational Psychology, 99,* 653–670. http://dx.doi.org/10.1037/0022-0663.99.3.653

Burke, M. A., & Sass, T. R. (2013). Classroom peer effects and student achievement. *Journal of Labor Economics, 31,* 51–82. http://dx.doi.org/10.1086/666653

Cabell, S. Q., DeCoster, J., LoCasale-Crouch, J., Hamre, B. K., & Pianta, R. C. (2013). Variation in the effectiveness of instructional interactions across preschool classroom settings and learning activities. *Early Childhood Research Quarterly, 28,* 820–830. http://dx.doi.org/10.1016/j.ecresq.2013.07.007

Carr, E. G., Taylor, J. C., & Robinson, S. (1991). The effects of severe behavior problems in children on the teaching behavior of adults. *Journal of Applied Behavior Analysis, 24,* 523–535. http://dx.doi.org/10.1901/jaba.1991.24-523

Chase-Lansdale, P. L., Moffitt, R. A., Lohman, B. J., Cherlin, A. J., Coley, R. L., Pittman, L. D., . . . Votruba-Drzal, E. (2003). Mothers' transitions from welfare to work and the well-being of preschoolers and adolescents. *Science, 299,* 1548–1552. http://dx.doi.org/10.1126/science.1076921

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Coleman, J. S., Campbell, E. Q., Hobson, C., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Government Printing Office.

Connor, C. M., Son, S.-H., Hindman, A. H., & Morrison, F. J. (2005). Teacher qualifications, classroom practices, family characteristics, and preschool experience: Complex effects on first graders' vocabulary and early reading outcomes. *Journal of School Psychology, 43,* 343–375. http://dx.doi.org/10.1016/j.jsp.2005.06.001

Cooley, J. (2010). *Can achievement peer effects estimates inform policy? A view from inside the black box.* Unpublished manuscript. Retrieved from http://www.ssc.wisc.edu/~jcooley/exogpeer.pdf

Crane, J. (1991). The epidemic theory of ghettos and neighborhood effects on dropping out and teenage childbearing. *American Journal of Sociology, 96,* 1226–1259. http://dx.doi.org/10.1086/229654

Creemers, B. P., & Reezigt, G. J. (1996). School level conditions affecting the effectiveness of instruction. *School Effectiveness and School Improvement, 7,* 197–228. http://dx.doi.org/10.1080/0924345960070301

Crosnoe, R. (2009). Students from low-income households and the socioeconomic composition of public high schools. *American Sociological Review, 74,* 709–730. http://dx.doi.org/10.1177/000312240907400502

Dishion, T. J., Spracklen, K. M., Andrews, D. W., & Patterson, G. R. (1996). Deviancy training in male adolescent friendships. *Behavior Therapy, 27,* 373–390.

Dotterer, A. M., Burchinal, M., Bryant, D., Early, D., & Pianta, R. C. (2013). Universal and targeted pre-kindergarten programmes: A comparison of classroom characteristics and child outcomes. *Early Child Development and Care, 183,* 931–950. http://dx.doi.org/10.1080/03004430.2012.698388

Dreeben, R., & Barr, R. (1988). Classroom composition and the design of instruction. *Sociology of Education, 61,* 129–142. http://dx.doi.org/10.2307/2112622

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43,* 1428–1446. http://dx.doi.org/10.1037/0012-1649.43.6.1428

Duncan, G. J., & Magnuson, K. (2011). The nature and impact of early achievement skills, attention skills, and behavior problems. In G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 47–69). New York, NY: Russell Sage Foundation.

Duncan, G. J., Ziol-Guest, K. M., & Kalil, A. (2010). Early-childhood poverty and adult attainment, behavior, and health. *Child Development, 81,* 306–325. http://dx.doi.org/10.1111/j.1467-8624.2009.01396.x

Duncan, S. E., & De Avila, E. A. (1998). *PreLAS 2000*. Monterey, CA: CTB/McGraw-Hill.

Dunn, L. M., & Dunn, L. M. (1997). *Examiners manual for the PPVT-III: Peabody Picture Vocabulary Test*. Circle Pines, MN: American Guidance Service.

Dunn, L. M., Padilla, E. R., Lugo, D. E., & Dunn, L. M. (1986). *Test de Vocabulario en Imagenes Peabody: TVIP*. Circle Pines, MN: American Guidance Service.

Early, D. M., Barbarin, O., Bryant, D., Burchinal, M., Chang, F., & Clifford, R. (2005). *Prekindergarten in eleven states: NCEDL's multi-state study of pre-kindergarten and study of state-wide early education programs (SWEEP)*. Chapel Hill, NC: University of North Carolina.

Early, D. M., Iruka, I. U., Ritchie, S., Barbarin, O. A., Winn, D. M. C., Crawford, G. M., . . . Pianta, R. C. (2010). How do pre-kindergarteners spend their time? Gender, ethnicity, and income as predictors of experiences in pre-kindergarten classrooms. *Early Childhood Research Quarterly, 25,* 177–193.

Gamoran, A. (1986). Instructional and institutional effects of ability grouping. *Sociology of Education, 59,* 185–198. http://dx.doi.org/10.2307/2112346

Garcia, E. (2015). *Inequality at the starting gate: Cognitive and noncognitive skills gaps between 2010–2011 kindergarten classmates*. Washington, DC: Economic Policy Institute.

Gelman, R., & Gallistel, C. R. (1986). *The child's understanding of number*. Cambridge, MA: Harvard University Press.

Georges, A., Brooks-Gunn, J., & Malone, L. M. (2012). Links between young children's behavior and achievement: The role of social class and classroom composition. *American Behavioral Scientist, 56,* 961–990. http://dx.doi.org/10.1177/0002764211409196

Ginsburg, H. P., Lee, J. S., & Boyd, J. S. (2008). Math education for young children: What it is and how to promote it. *Social Policy Report: Giving Child and Youth Development Knowledge Away, 22,* 3–23.

Gordon, R. A., & Chase-Lansdale, P. L. (2001). Availability of child care in the United States: A description and analysis of data sources. *Demography, 38,* 299–316.

Gormley, W. T., Jr., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology, 41,* 872–884. http://dx.doi.org/10.1037/0012-1649.41.6.872

Gormley, W. T., Jr., Phillips, D., & Gayer, T. (2008). The early years: Preschool programs can boost school readiness. *Science, 320,* 1723–1724. http://dx.doi.org/10.1126/science.1156019

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics, 18,* 527–544. http://dx.doi.org/10.1002/jae.741

Harwell, M., & LeBeau, B. (2010). Student eligibility for a free lunch as an SES measure in education research. *Educational Researcher, 39,* 120–131. http://dx.doi.org/10.3102/0013189X10362578

Hastie, T., & Tibshirani, R. (1990). *General additive models*. New York, NY: Chapman Hall.

Henry, G. T., Henderson, L. W., Ponder, B. D., Gordon, C. S., Mashburn, A. J., & Rickman, D. K. (2003). *Report of the findings from the Early Childhood Study: 2001–2002*. Atlanta, GA: Georgia State University.

Henry, G. T., & Rickman, D. K. (2007). Do peers influence children's skill development in preschool? *Economics of Education Review, 26,* 100–112.

Hinshaw, S. P. (1992). Academic underachievement, attention deficits, and aggression: Comorbidity and implications for intervention. *Journal of Consulting and Clinical Psychology, 60,* 893.

Hoxby, C. M., & Weingarth, G. (2005). *Taking race out of the equation: School reassignment and the structure of peer effects*. Unpublished manuscript, Harvard University, Cambridge, MA.

Hutchison, D. (2003). The effect of group-level influences on pupils' progress in reading. *British Educational Research Journal, 29,* 25–40. http://dx.doi.org/10.1080/0141192032000057357

Imberman, S. A., Kugler, A. D., & Sacerdote, B. I. (2012). Katrina's children: Evidence on the structure of peer effects from hurricane evacuees. *The American Economic Review, 102,* 2048–2082. http://dx.doi.org/10.1257/aer.102.5.2048

Johnson, M. P., Ladd, H. F., & Ludwig, J. (2002). The benefits and costs of residential mobility programmes for the poor. *Housing Studies, 17,* 125–138. http://dx.doi.org/10.1080/02673030120105947

Justice, L. M., Petscher, Y., Schatschneider, C., & Mashburn, A. (2011). Peer effects in preschool classrooms: Is children's language growth associated with their classmates' skills? *Child Development, 82,* 1768–1777. http://dx.doi.org/10.1111/j.1467-8624.2011.01665.x

Lavy, V., Paserman, M. D., & Schlosser, A. (2012). Inside the black of box of ability peer effects: Evidence from variation in low achievers in the classroom. *The Economic Journal, 122,* 208–237. http://dx.doi.org/10.1111/j.1468-0297.2011.02463.x

Lazear, E. P. (2001). Educational production. *The Quarterly Journal of Economics, 116,* 777–803. http://dx.doi.org/10.1162/00335530152466232

LeFevre, J. A., Skwarchuk, S. L., Smith-Chant, B. L., Fast, L., Kamawar, D., & Bisanz, J. (2009). Home numeracy experiences and children's math performance in the early school years. *Canadian Journal of Behavioural Science, 41,* 55.

LoCasale-Crouch, J., Konold, T., Pianta, R., Howes, C., Burchinal, M., Bryant, D., . . . Barbarin, O. (2007). Profiles of observed classroom quality in state-funded pre-kindergarten programs and associations with teacher, program, and classroom characteristics. *Early Childhood Research Quarterly, 22,* 3–17. http://dx.doi.org/10.1016/j.ecresq.2006.05.001

Magnuson, K. A., Meyers, M. K., Ruhm, C. J., & Waldfogel, J. (2004). Inequality in preschool education and school readiness. *American Educational Research Journal, 41,* 115–157. http://dx.doi.org/10.3102/00028312041001115

Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review, 26,* 33–51. http://dx.doi.org/10.1016/j.econedurev.2005.09.008

Magnuson, K. A., & Votruba-Drzal, E. (2009). Enduring influences of childhood poverty. In M. Cancian & S. Danziger (Eds.), *Changing Poverty, Changing Policies* (pp. 153–179). New York, NY: Russell Sage Foundation.

Magnuson, K. A., & Waldfogel, J. (2005). Early childhood care and education: Effects on ethnic and racial gaps in school readiness. *The Future of Children, 15,* 169–196. http://dx.doi.org/10.1353/foc.2005.0005

Marsh, H. W., & Hau, K. T. (2003). Big-fish-little-pond effect on academic self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist, 58,* 364–376. http://dx.doi.org/10.1037/0003-066X.58.5.364

Mashburn, A. J., Justice, L. M., Downer, J. T., & Pianta, R. C. (2009). Peer effects on children's language achievement during pre-kindergarten. *Child Development, 80,* 686–702. http://dx.doi.org/10.1111/j.1467-8624.2009.01291.x

Mashburn, A., Pianta, R., Hamre, B., Downer, J., Barbarin, O., Bryant, D., . . . Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills.

Child Development, 79, 732–749. http://dx.doi.org/10.1111/j.1467-8624.2008.01154.x

National Center for Early Development and Learning. (2005). *Multi-state study of pre-kindergarten: Study description & documentation for public release of data.* Chapel Hill, NC: Author.

Neidell, M., & Waldfogel, J. (2010). Cognitive and noncognitive peer effects in early education. *The Review of Economics and Statistics, 92,* 562–576. http://dx.doi.org/10.1162/REST_a_00012

Office of the Press Secretary. (2013). Fact Sheet: President Obama's Plan for Early Education for all Americans. Retrieved from https://www.whitehouse.gov/the-press-office/2013/02/13/fact-sheet-president-obama-s-plan-early-education-all-americans

Pallas, A. M., Entwisle, D. R., Alexander, K. L., & Stluka, M. F. (1994). Ability-group effects: Instructional, social, or institutional? *Sociology of Education, 67,* 27–46. http://dx.doi.org/10.2307/2112748

Parrish, D., & Howes, C. (2008). *First 5 LA school readiness initiative parent and child outcomes study.* Washington, DC: American Institutes of Research.

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata.* College Station, TX: STATA Press.

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Bub, K., & Pressler, E. (2011). CSRP's impact on low-income preschoolers' preacademic skills: Self-regulation as a mediating mechanism. *Child Development, 82,* 362–378. http://dx.doi.org/10.1111/j.1467-8624.2010.01561.x

Reid, J. L., & Ready, D. D. (2013). High-quality preschool: The socioeconomic composition of preschool classrooms and children's learning. *Early Education and Development, 24,* 1082–1111. http://dx.doi.org/10.1080/10409289.2012.757519

Reynolds, A. J., & Temple, J. A. (1995). Quasi-experimental estimates of the effects of a preschool intervention: Psychometric and econometric comparisons. *Evaluation Review, 19,* 347–373. http://dx.doi.org/10.1177/0193841X9501900401

Rivkin, S. G. (2001). Tiebout sorting, aggregation and the estimation of peer group effects. *Economics of Education Review, 20,* 201–209. http://dx.doi.org/10.1016/S0272-7757(00)00032-7

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal, 4,* 227–241.

Schechter, C., & Bye, B. (2007). Preliminary evidence for the impact of mixed-income preschools on children from low-income households's language growth. *Early Childhood Research Quarterly, 22,* 137–146. http://dx.doi.org/10.1016/j.ecresq.2006.11.005

Setodji, C. M., Scheuner, M., Pankow, J. S., Blumenthal, R. S., Chen, H., & Keeler, E. (2012). A graphical method for assessing risk factor threshold values using the generalized additive model: The multi-ethnic study of atherosclerosis. *Health Services and Outcomes Research Methodology, 12,* 62–79. http://dx.doi.org/10.1007/s10742-012-0082-1

Stipek, D. (2004). Teaching practices in kindergarten and first grade: Different strokes for different folks. *Early Childhood Research Quarterly, 19,* 548–568. http://dx.doi.org/10.1016/j.ecresq.2004.10.010

Strand, S. (1998). A value added analysis of the 1996 primary school performance tables. *Educational Research, 40,* 123–137. http://dx.doi.org/10.1080/0013188980400201

Suls, J. (1986). Comparison processes in relative deprivation: A life-span analysis. In J. M. Olsen, C. P. Herman, & M. P. Zanna (Eds.), *Relative deprivation and social comparison: The Ontario Symposium* (Vol. 4, pp. 95–116). Hillside, NJ: Erlbaum.

Teddlie, C., Stringfield, S., & Reynolds, D. (2000). Context issues within school effectiveness research. In C. Teddlie & D. Reynolds (Eds.), *International handbook of school effectiveness research* (pp. 160–185). London, UK: Falmer Press.

Tudge, J. R., & Doucet, F. (2004). Early mathematical experiences: Observing young Black and White children's everyday activities. *Early Childhood Research Quarterly, 19,* 21–39. http://dx.doi.org/10.1016/j.ecresq.2004.01.007

U.S. Department of Health and Human Services, Administration for Children and Families, Office of Head Start. (2013). *Head Start Program facts: Fiscal year 2011.* Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/mr/factsheets/2011-hs-program-factsheet.html

Van Ewijk, R., & Sleegers, P. (2010). The effect of peer socioeconomic status on student achievement: A meta-analysis. *Educational Research Review, 5,* 134–150. http://dx.doi.org/10.1016/j.edurev.2010.02.001

Votruba-Drzal, E., Coley, R. L., Koury, A. S., & Miller, P. (2013). Center-based child care and cognitive skills development: Importance of timing and household resources. *Journal of Educational Psychology, 105,* 821–838. http://dx.doi.org/10.1037/a0032951

Weigel, D. J., Martin, S. S., & Bennett, K. K. (2006). Mothers' literacy beliefs: Connections with the home literacy environment and pre-school children's literacy development. *Journal of Early Childhood Literacy, 6,* 191–211. http://dx.doi.org/10.1177/1468798406066444

Weiland, C. (2013, April). *Does peer socioeconomic status in preschool predict children's executive function and language skills?* Paper presented at the Biennial Meeting of the Society for Research in Child Development, Seattle, WA.

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development, 84,* 2112–2130. http://dx.doi.org/10.1111/cdev.12099

Whitehurst, G. J., & Lonigan, C. J. (1998). Emergent literacy: Development from prereaders to readers. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (pp. 11–29). New York, NY: Guilford Press.

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management, 27,* 122–154. http://dx.doi.org/10.1002/pam.20310

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson tests of achievement.* Itasca, IL: Riverside Publishing.

Woodcock, R. W., & Sandoval, A. F. M. (1996). *Batería Woodcock-Muñoz: Pruebas de aprovechamiento revisada.* Itasca, IL: Riverside Publishing.

Zill, N., Resnick, G., Kim, K., O'Donnell, K., Sorongon, A., McKey, R. H., & D'Elio, M. A. (2003). *Head Start FACES 2000: A whole-child perspective on program performance. Fourth progress report.* Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families.

Zimmer, R. W., & Toma, E. F. (2000). Peer effects in private and public schools across countries. *Journal of Policy Analysis and Management, 19,* 75–92.

Received October 16, 2014
Revision received April 27, 2016
Accepted May 1, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 2, 166–177

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000139

School Transition Practices and Children's Social and Academic Adjustment in Kindergarten

Kyle DeMeo Cook and Rebekah Levine Coley
Boston College

The transition to kindergarten is a critical period for children and families, with successful transitions setting the stage for short- and long-term academic and social success. This study explored the practices used by kindergarten teachers to help ease children's and families' transition into primary school (termed "transition practices"), and assessed their relationship to children's social and academic adjustment to school in a nationally representative sample of children in the United States ($N = 4,900$). On average, kindergarten teachers engaged in 3 transition practices, with outreach to parents and child or parent classroom visits most common, and structural changes to the school schedule less frequent. Private schools and more experienced teachers engaged in more transition practices, whereas ethnic and racial minority, immigrant, and urban children had teachers who reported fewer practices. Prospective, lagged regression models found that engagement in more types of transition practices was predictive of heightened prosocial behaviors among children, but was not associated with children's attention or academic outcomes. Examination of specific types of practices found that transition activities geared toward parents were associated with children's heightened academic skills in kindergarten. These results provide limited evidence to support the "more is better" view of transition practices and instead suggest that specific types of transition practices are linked to particular aspects of children's functioning.

Keywords: kindergarten, school readiness, school transition, adjustment, family school connections

Supplemental materials: http://dx.doi.org/10.1037/edu0000139.supp

The transition to school is an important time in the lives of children and families. When children enter kindergarten, they often face a qualitatively different environment than their homes and previous early education and care programs, with increased demands and expectations for both parents and children (Cowan, Cowan, Ablow, Johnson, & Measelle, 2005; Pianta, Cox, & Snow, 2007; Pianta & Kraft-Sayre, 2003; Ramey & Ramey, 2010; Rimm-Kaufman & Pianta, 2000). In a national survey on the transition to kindergarten, teachers reported that almost half (48%) of children had some difficulty adjusting to school, with 16% having serious difficulties and 32% having some difficulties (Rimm-Kaufman, Pianta, & Cox, 2000). School entry marks a transition not just for children but also their parents, with shifting identities and decreasing opportunities to engage in their child's day to day activities (Cowan et al., 2005).

The prevalence of difficulties adjusting to school is important, given that successful transitions provide children with the foundation for later school success. Entering school is a critical period of cognitive and social development for children, a time when many basic and foundational skills are taught and often when student records begin that may follow the child for the duration of their school experience (Entwisle & Alexander, 1993). During this period, "achievement trajectories" are launched with positive early school experiences and adjustment having implications for success at school entry and beyond (Entwisle & Alexander, 1993; Snow, 2006). Research supports that early school experiences are predictive of later school achievement (Pianta & Walsh, 1996; Reynolds, 2004), with children who have positive experiences more likely to report enjoying school and having fewer absences, thus potentially gaining more from the available academic experiences that lead to better academic and social outcomes (Ladd, Buhs, & Seid, 2000; Ladd & Price, 1987; Pianta & Kraft-Sayre, 2003).

A Developmental Ecological View of Transitions

Given the importance of smooth and positive transitions, it is essential to better understand correlates of successful transitions to kindergarten and to delineate whether practices undertaken by schools to support this transition are associated with more successful immediate and long-term functioning for children. Rimm-Kaufman and Pianta's (2000) developmental ecological transition to kindergarten model emphasizes that successful transitions are embedded within interacting systems and rely on connections between families, early childhood settings, and the elementary schools children are entering. The National Education Goals Panel

This article was published Online First June 20, 2016.

Kyle DeMeo Cook and Rebekah Levine Coley, Applied Developmental & Educational Psychology, Boston College.

We gratefully acknowledge the support of the Spencer Foundation (201300115). The content is solely the responsibility of the authors and does not necessarily represent the official views of the grantors. A special thank you is also extended to the children and families that participated in the Early Childhood Longitudinal Study, Birth Cohort.

Correspondence concerning this article should be addressed to Kyle DeMeo Cook, Applied Developmental & Educational Psychology, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467. E-mail: demeok@bc.edu

of 2000, which called for all children to start school "ready to learn," exemplifies the tenets of the ecological and dynamic view of the transition to school, arguing that in order to have children ready for school, there is also a need for "ready schools," "ready families," and "ready communities" (National Education Goals Panel, 1998). Using this model, "readiness" becomes the property of all the pieces of the system, not just the child (Pianta & Kraft-Sayre, 2003).

Ramey, Ramey, and Lanzi (2006) suggest that there are multiple components that contribute to successful school transitions, including that children and parents have positive attitudes toward school and learning, parents and key adults act as partners in children's learning, and teachers value children as individuals and provide developmentally appropriate early experiences. Although the literature suggests that successful transitions need to be the responsibility of all parties, including parents, early education providers, kindergarten teachers, schools, and other service providers (Kagan & Neuman, 1998), the onus to provide support generally falls to the elementary schools children are entering.

Description of School-Based Transition Practices

Transition practices implemented by schools can help serve as a bridge for children and families as they move into kindergarten. Within the developmental ecological model, successful transition activities foster positive relationships and should include connections between children and families and schools (Pianta & Kraft-Sayre, 2003). Activities specifically targeted at children may include visiting the kindergarten classroom, meeting their new teacher, and learning about what to expect from school. Activities specifically targeted at families may include attending conferences, registration, and open houses, or familiarizing themselves with school practices and policies through written materials. Some of these practices are fairly common, such as schools sending information home to families, organizing parent orientations, and having open houses, whereas other more intensive practices—such as teachers conducting home visits or preschool visits—are much less common (Pianta, Cox, Taylor, & Early, 1999; Schulting, Malone, & Dodge, 2005).

Past research has shown that most school districts engage in some transition activities, and when offered, most families participate and find them helpful (La Paro, Kraft-Sayre, & Pianta, 2003). Yet there is limited research delineating which characteristics of schools, teachers, families, and children are associated with the use of transition practices. For example, teachers with bigger classes and less training on transitions may engage in fewer transition practices (Early, Pianta, & Cox, 1999). In addition, schools that serve more high-poverty and minority children have been found to engage in fewer transition practices overall, and specifically in less individualized practices aimed at families (Early et al., 1999; Love, Logue, Trudeau, & Thayer, 1992; Pianta et al., 1999; Schulting et al., 2005). This is particularly troubling because teachers in schools with a high composition of minority students and higher district poverty levels report increased difficulties in students' adjustment to school compared with teachers from schools serving more advantaged children (Rimm-Kaufman & Pianta, 2000). Given the limited knowledge in this arena, more comprehensive research is needed to further identify the child, family, teacher, and school characteristics associated with greater and lesser engagement in transition practices.

Evidence Relating Transition Practices to Children's Adjustment to School

Beyond the need for more descriptive information, a central question is whether transition practices help improve child experiences in school. Further, it is important to address which specific types of transition practices are most effective, and for whom. Although theorists and educators have increasingly maintained the importance of school transition practices, there is limited empirical evidence to support their effectiveness in easing children's entry into kindergarten or improving their cognitive and behavioral success after starting school. A review of the research on transition practices for typically developing children (Eckert et al., 2008) found only a few published studies, with only one study explicitly exploring the relationship between the transition practices used by kindergarten teachers and children's academic outcomes (Schulting et al., 2005). Schulting and colleagues (2005) assessed transition practices addressing information targeted to parents (e.g., the school telephoned or sent information home, the teacher visited the home, the parents visited the school, or there was a parent orientation), practices targeting the child (e.g., the child visited the classroom, or the school had shortened days for new kindergarteners), and an open "other" category. Using data on a nationally representative sample of over 17,000 children from the Early Childhood Longitudinal Study, Kindergarten Cohort of 1998, Schulting and colleagues summed the number of school-based transition practices reported by kindergarten teachers into a total score, finding that a greater number of transition practices was associated with heightened academic achievement scores among children at the end of the school year.

In a study of about 400 children transitioning to formal schooling in Finland, preschool and elementary school teacher pairs reported on the practices they collaborated on during children's preschool year to aid in the transition. They found that a greater number of transition activities was associated with heightened growth in children's reading, writing, and math skills through the first year of elementary school (Ahtola et al., 2011). These results are consistent with and extend Schulting and colleagues' (2005) results by assessing growth in children's outcomes. However, both of these studies focused only on academic outcomes, and did not consider associations between transition practices and other key aspects of successful school transitions, such as children's initial adjustment to school or longer term social and behavioral functioning through kindergarten.

Additional research has assessed the association between transition practices by preschool teachers and child functioning in kindergarten. LoCasale-Crouch, Mashburn, Downer, and Pianta (2008) studied the transition to school for a sample of approximately 320 children attending public prekindergarten programs in the United States, finding a positive link between the number of transition activities used by preschool teachers (e.g., children, teachers, and/or parents visiting kindergarten, sharing information between preschool and kindergarten teachers) and children's social, self-regulation, and academic skills in the fall of their kindergarten year.

In addition to assessing whether the breadth of transition practices was supportive of heightened child functioning, these studies also assessed whether specific individual activities were more or less important in promoting children's success in kindergarten. For example, Schulting and colleagues (2005) found that parents and children visiting the kindergarten classroom before the school year started were the only individual transition practices to significantly predict heightened academic scores among children (Schulting et al., 2005). These results reiterate arguments from other research suggesting that collaboration between schools and families improves children's school success in realms including more positive attitudes toward school, better attendance, higher grades, and higher graduation rates (Dearing, Kreider, & Weiss, 2008; Henderson & Berla, 1994; Pomerantz, Moorman, & Litwack, 2007). In contrast, the studies that explored activities engaged in by preschool teachers found that sharing information with kindergarten teachers about curriculum or individual children were the activities most predictive of better child outcomes (Ahtola et al., 2011; LoCasale-Crouch et al., 2008).

A final central question in the literature concerns individual differences among children, that is, asking whether kindergarten transition practices are more important for certain children than for others. For example, Schulting and colleagues (2005) hypothesized that transition practices might be most protective for children facing greater risks of poor school adjustment because of limited family economic resources. Consistent with past descriptive research on transition practices, this study reported that children from families with low socioeconomic status received the fewest number of transition practices, yet gained the most from them. LoCasale-Crouch and colleagues (2008) found similar results when exploring whether transition practices moderated the relationship between risk factors (i.e., maternal education, poverty level, and race) and child outcomes. They found that transition practices were more strongly related to children's functioning for children facing risk factors than for their more advantaged peers.

Research Goals

The present study sought to expand the limited literature in this arena by assessing the transition to kindergarten in a large, nationally representative sample of American children born in 2001 and followed from infancy through the transition to kindergarten in 2006 or 2007. The use of longitudinal data allowed us to assess children's family characteristics, early educational experiences, and functioning prior to kindergarten entry, important information for delineating the populations most and least likely to experience transition practices, and also essential for helping to isolate unique associations between transition practices and children's emotional, behavioral, and academic functioning after starting kindergarten. This is an important contribution to the literature, as past studies on U.S. samples did not account for children's earlier functioning before school entry.

One set of goals of the current study was descriptive. First, we sought to provide an updated profile of kindergarten transition practices, exploiting the generalizability of a nationally representative sample and considering a broad set of transition practices engaged in by kindergarten teachers. Based on prior research (Pianta et al., 1999; Schulting et al., 2005), we expected that practices providing information to parents and organizing child classroom visits would be most common. Second, we sought to provide a rich description of child, parent, teacher, and school characteristics associated with teachers' engagement in transition practices. We expected that less experienced teachers and those serving more disadvantaged children would report fewer transition practices, although it was difficult to develop further hypotheses concerning child and family characteristics from the limited research base.

The third and most substantive goal of this study was to extend evidence on connections between transition practices and children's successful kindergarten functioning. In particular, we sought to extend prior literature by (a) considering not only children's academic skills but also their behavioral adjustment and social skills in kindergarten; (b) adjusting for children's prior functioning as well as child, family, and early childhood education (ECE) experiences in order to isolate associations between transition practices and children's growth in functioning; and (c) assessing whether a "more is better" model best explains links between transition practices and children's functioning, or rather whether certain types of transition practices appear most effective at supporting particular aspects of children's success in kindergarten. We expected that more practices would predict better child outcomes, and that parent and child visits to the kindergarten class would be the most important individual practices. In addition, following tenets of the developmental ecological transition to kindergarten model (Rimm-Kaufman & Pianta, 2000) and evidence that transition practices are more important for at-risk children (LoCasale-Crouch et al., 2008; Schulting et al., 2005), we assessed whether family income moderated associations between transition practices and children's functioning. We expected that transition practices helping to familiarize children and families with the norms and practices of kindergarten would be particularly important for the functioning of children from low-income families.

Method

Participants

Data were drawn from the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), a longitudinal multicomponent study following a nationally representative sample of approximately 10,700 children (the ECLS-B requires that all *N*s be rounded to the nearest 50) born in the United States in 2001 from infancy through kindergarten entry (Chernoff, Flanagan, McPhee, & Park, 2007). Children who died or were adopted prior to 9 months of age and children born to mothers under 15 years of age were excluded from the sample. The ECLS-B collected five or six waves of data from primary caregiver interviews (with the child's mother in 98% of cases) and child assessments when children were (on average) 10 months, 2 years, 4 years, 5 years, and, for the approximately 25% of children not yet in kindergarten by Age 5, 6 years of age. The response rate was 74% at the first wave, followed by rates of 93%, 91%, 92%, and 92% among children remaining in the sample at each wave. Data for this study were drawn from children's kindergarten year (Wave 4 or 5) and the year prior to kindergarten (Wave 3 or 4). Kindergarten teachers were interviewed in Waves 4 or 5 (response rates of 74% and 76%), and ECE providers were interviewed for children in regular center or home-based ECE settings in Waves 3 or 4 (response rates of 70% and 87%).

Because we are focused on transition practices led by kindergarten teachers and children's functioning in kindergarten, the analytic sample focused on children who remained in the sample at kindergarten and had kindergarten teacher interview data, approximately 5,050 children. Of these children, about 150 were excluded because they were not first-time kindergarteners or did not attend a kindergarten classroom at school entry (e.g., went directly to first grade or were in ungraded classroom). This resulted in approximately 4,900 children in the study's analytic sample. Children in the sample were 51% White and 51% male. They averaged 68 months old at kindergarten entry, 89% attended public schools, and 75% were enrolled in full-day kindergarten classrooms. At the kindergarten wave, there was substantial geographic dispersion of the sample, resulting in approximately one study child per school (Snow et al., 2009). It is essential to note that the ECLS-B calculated weights that adjust for differential sampling and nonresponse, as well as attrition over the waves. To adjust for these factors and properly estimate standard errors, given the complex sampling design, all analyses included 90 replicate weights (wk45t1-wk45t90) using jackknife replication methods as suggested by the ECLS-B (Snow et al., 2009). This set of weights was carefully chosen from all weights created by the ECLS-B to fit our exact analytic sample (children starting kindergarten in 2006 or 2007 with parent and kindergarten teacher interview data), adjusting for teacher nonresponse as well as child and parent attrition through the waves. The use of these weights allows us to generalize results to all children born in the United States in 2001.

Prior to conducting analyses, we explored the presence of missing data in the analytic sample of approximately 4,900 children. Item-level missing data ranged from 0% to 20%. Little's missing completely at random (MCAR) test was performed in SPSS 22, and revealed that missing values in the analytic sample were not MCAR, $\chi^2 = 2{,}333.72$ (1064), $p = .00$. Further, observed variables were related to missingness (with a general pattern suggesting that children with lower functioning and few social and economic resources were more likely than their peers to have missing data), supporting the appropriateness of multiple imputation to address missing data (Little, 1988). Missing data were imputed using multiple imputation by chained equations in Stata 12.1 (Royston, 2005) to create 20 complete data sets. The imputation models included all variables described in the measures section, incorporating ordinary least squares (OLS), logit, ordered logit, multinomial logit, and Poisson modeling techniques as appropriate, depending upon the scaling of the variables. Following imputation, analyses were run using the mi estimate command in Stata in order to aggregate results and properly estimate standard errors across the imputed data sets.

Measures

Children's behavioral school adjustment. Children's behavioral adjustment and functioning in kindergarten were assessed via parent and kindergarten teacher reports. Parents reported on children's short-term adjustment to kindergarten through six items assessing children's behaviors in the first 2 weeks after school entry. Items rated how often the child complained about school, was reluctant to go to school, pretended to be sick, said good things about school, reported liking school, and looked forward to going to school, on a scale from 1 to 3, with items recoded so that higher scores showed more positive adjustment to school (Chernoff et al., 2007). Factor analysis, conducted in STATA using a polychoric correlation matrix to account for the ordinal nature of the variables, revealed that items loaded onto one factor, and, thus, items were averaged into an adjustment scale ($\alpha = .69$). Teachers reported on a broader set of children's behaviors, with ratings completed, on average, 2.3 months after the start of school. Teachers reported on a series of items drawn from well-validated measures, including the Preschool and Kindergarten Behavior Scales—Second Edition (Merrell, 2003), the Social Skills Rating Scales (Gresham, Elliott, & Black, 1987), and the Family and Child Experiences Study. Teachers rated the frequency of the child's engagement in behaviors on 5-point scales (*never* to *very often*). We used composite measures validated in prior research with the ECLS-B (Coley, Votruba-Drzal, Collins, & Cook, 2016; Coley, Votruba-Drzal, Miller, & Koury, 2013), and reestablished by factor analyses in Stata using polychroic correlation matrices, to assess children's prosocial and attention skills. Prosocial skills consisted of an average of six items assessing behaviors such as making friends, sharing, and comforting others ($\alpha = .82$). Attention skills were assessed with five items delineating children's attention, independence, task completion, and eagerness to learn ($\alpha = .83$).

Children's academic adjustment. Children's cognitive skills were assessed after kindergarten entry (2.3 months, on average, after kindergarten entry) through direct assessments. Assessments incorporated items drawn from well-validated, standardized instruments such as the Peabody Picture Vocabulary Test—Third Edition (L. M. Dunn & Dunn, 1997), the PreLAS 2000 (Duncan & DeAvila, 1998), the Preschool Comprehensive Test of Phonological & Print Processing (Lonigan, Wagner, Torgeson, & Rashotte, 2002), and the Test of Early Mathematics Ability (3rd ed.; Ginsburg & Baroody, 2003). ECLS-B statisticians completed extensive data cleaning and validation work on these cognitive measures, using Item Response Theory methods to create composites, which we used in our analyses (see Snow et al., 2009, for details). The early reading assessment ($\alpha = .92$) consisted of 74 items that measured early reading and language skills, including letter knowledge, word recognition, print conventions, and phonological awareness. The math assessment ($\alpha = .92$) consisted of 58 items focused on number sense, properties, operations, and probability. Children's kindergarten behavioral and academic adjustment measures were used as dependent variables in the child outcome models.

Reports of children's prosocial and attention skills and direct assessments of children's cognitive skills were also collected at the preschool wave. Following prior research (e.g., Coley et al., 2016) preschool ratings of children's prosocial and attention skills were reported by the provider of their main educational environment. For seventy seven percent of the sample, early education providers reported children's behaviors (prosocial skills, six items, $\alpha = .81$; attention skills, five items, $\alpha = .83$). For the remaining 23% who were not in an early education setting at Age 4, parents reported on child behaviors (prosocial skills, six items, $\alpha = .80$; attention skills, four items [one item specifically addressing attention in school was not reported by parents]. $\alpha = .65$). Children's early reading ($\alpha = .84$) and mathematics ($\alpha = .89$) skills were directly assessed using the same measures described for the kindergarten wave. Preschool cognitive and behavioral measures were used as

independent variables in the models predicting kindergarten transition practices and as covariates (lags of the dependent variable) to adjust for prior functioning in the child outcome models predicting kindergarten adjustment.

School transition practices. In the teacher questionnaire, kindergarten teachers reported whether they or others in their school engaged in seven different specific practices to make the transition to kindergarten less difficult for children in the study child's class. These included (a) phone/send home information about the kindergarten program to the parents; (b) invite parents to the school for orientation prior to the start of the school year; (c) have preschoolers spend some time in the kindergarten classroom prior to school; (d) have parents and children visit kindergarten prior to the start of the school year; (e) conduct home visits to the homes of children at the beginning of the school year; (f) shorten school days at the beginning of the school year for kindergarteners; and (g) stagger school entry so that kindergarteners start the school year in smaller groups before meeting with the full class. Each item was scored yes or no. Individual items were used as separate indicators and, following prior research (LoCasale-Crouch et al., 2008; Schulting et al., 2005), also were summed into a total activities index variable to delineate the breadth of transition activities engaged in by the teacher. It is important to note that teachers reported on the activities they engaged in; the ECLS-B did not gather data on whether individual children and parents participated in the offered activities. The full summed activity index was used as the dependent variable in the first model, with child, family, and school characteristics predicting transition practices. The summed activity index and individual items were also used as the main independent variables of interest in the models predicting child outcomes.

Child and family characteristics. The ECLS-B assessed a rich set of child and family characteristics, which were used both as predictors of transition practices and as covariates in models predicting children's successful transition to kindergarten. These include indicator variables of children's male gender, whether they were part of a multiple child birth, whether they were born with low birth weight (less than 2,500 g/5.5 pounds), and whether they were ever diagnosed with a cognitive, behavioral, or physical disability prior to kindergarten (all reported by parents and coded as 0 or 1). Children's preschool type was reported by parents and teachers and coded into mutually exclusive categories of no nonparental preschool experience, home-based in child's home, home-based in another home, center-based in a school, and center-based in another location. Parental race and ethnicity and immigrant status were combined into a set of mutually exclusive groups delineating children of native-born Whites, native-born Blacks, native-born Hispanics, Native American Indians, and native-born children of other races (including parents of different races, and "other"), Hispanic immigrants, Asians (94% of whom were immigrants), and other immigrants. An additional indicator designated families whose primary language was not English. Mother's age at first childbirth was also reported as well as highest level of parental education, which was coded into mutually exclusive groups (less than a high school diploma, high school diploma or GED, some college or vocational training, or bachelor's degree or higher). Each family's community was coded as urban, suburban, or rural.

Other characteristics of parents and families that may shift over time were assessed at the preschool wave, the wave just prior to each child's kindergarten entry. These included indicators of maternal employment, maternal marital status, and maternal depression (assessed using a modified version of the Center for Epidemiological Studies Depression Scale [Radloff, 1977; 12 items, α = .89], dichotomized to designate whether or not scores were in the moderately to severely depressed range), all coded 0 or 1, as well as continuous variables of total family income (in units of \$10,000), the number of nonparental adults in the household, and the number of children in the household.

Because instability in children's home environments may affect their successful transition to kindergarten (Cowan et al., 2005), we also considered each of these measures assessed at the kindergarten wave, coded to indicate whether or not a transition had occurred (e.g., a transition from married to single or from single to married; a loss or gain in the number of children in the household). Because transitions in and of themselves, whether they are commonly seen as "positive" or "negative," all cause disequilibrium and require adjustment in order for healthy development to occur (Cowan & Heming, 2005; Erikson, 1950), we coded change in either direction as "1," with stability coded "0." Following prior research, which has found that aspects of instability tend to co-occur (Kull, Coley, & Lynch, 2015; Vernon-Feagans, Garrett-Peters, Willoughby, Mills-Koonce, & the Family Life Project Key Investigators, 2012), indicators of instability were summed into a family instability index.

Kindergarten teacher and school characteristics. Kindergarten teachers reported on a number of school and classroom characteristics, including whether the school was public or private, whether the study child was attending a full-day or half-day program, the size of the class, and their years teaching. Parents reported whether the study child had siblings in the same school and the distance from their home to the child's school (less than 1 mile, 1–2.5 miles, 2.6–5 miles, 5.1–10 miles, or 10 or more miles).

The child, family, and school characteristics were chosen based on theory and research supporting their relation to early school success (Ahtola et al., 2011; Cowan et al., 2005; LoCasale-Crouch et al., 2008; Rimm-Kaufman & Pianta, 2000; Schulting et al., 2005). Prior to inclusion, bivariate correlations were estimated to explore whether each covariate was associated with both transition practices and the child outcomes. All covariates included in the models were significantly correlated with one or more transition practices and one or more child outcome measures (see the online supplemental materials).

Analytic Plan

The first research aim of this study was to gain a better understanding of the prevalence of kindergarten transition practices, which we address through descriptive statistics. The second aim, assessing how child, family, and school characteristics were associated with transition practices, was addressed using a Poisson regression model in which all of the child, family, and school characteristics were used as predictors of the transition practices index score. These predictors included the four measures of children's functioning in the preschool wave (reading skills, math skills, prosocial behaviors, and attention skills). To address concerns over multicollinearity and for the sake of parsimony, the reading and math scores were averaged into one cognitive skills composite (r = .79), and the prosocial and attention skills measures were averaged into one behavioral composite (r = .59) for this analysis. As a robustness check, the model was also run with the individual functioning variables as predictors, and the results did not change. The more parsimonious model is presented in Table 1.

Table 1

Descriptive Statistics and Poisson Regression Model Predicting School Transition Practices From Child, Family, School, and Teacher Characteristics

Variables	Descriptive statistics		Poisson regression model predicting transition practices sum index	
	Mean or %	(*SD*)	IRR/*B*	(*SE*)
Transition practices sum index	3.29	(1.14)	—	—
Child and family characteristics				
Preschool wave cognitive functioning (math and reading)	29.15	(9.64)	1.01/.01	(.01)
Preschool wave behavioral functioning (attention and prosocial skills)	3.88	(.59)	1.00/.00	(.01)
Child disability	15%	—	1.05/.05**	(.02)
Male	51%	—	.99/–.01	(.01)
Twin	3%	—	.99/–.01	(.02)
Low birth weight	8%	—	.97/–.03+	(.02)
In kindergarten in 2007	27%	—	1.00/.00	(.03)
Age at kindergarten entry (months)	68.13	(4.39)	1.00/.00	(.00)
Months in kindergarten	2.31	(1.40)	1.01/.01*	(.01)
Siblings in same elementary school	40%	—	.98/–.02	(.02)
No ECE	23%	—	—	—
Home ECE in child's home	5%	—	1.03/.03	(.04)
Home ECE in other home	12%	—	1.02/.02	(.03)
Center ECE in school	18%	—	1.01/.01	(.03)
Center ECE in other	42%	—	1.01/.01	(.02)
Native White	51%	—	—	
Native Black	12%	—	.93/–.07**	(.02)
Native Hispanic	10%	—	.94/–.06*	(.03)
Native American	1%	—	.92/–.08	(.05)
Native multiple races	3%	—	.94/–.06	(.03)
Asian	3%	—	.99/–.01	(.03)
Non-Hispanic immigrant	5%	—	.92/–.09*	(.03)
Hispanic immigrant	15%	—	.89/–.12**	(.04)
Non-English household	18%	—	.94/–.06	(.03)
Family income ($10,000s)	5.89	(4.91)	1.00/.00	(.00)
Maternal employment	59%	—	1.00/.00	(.02)
Maternal depression	15%	—	.97/–.03	(.02)
Married parents	67%	—	1.00/.00	(.02)
Additional adults in household	30%	—	1.00/.00	(.02)
Children in household	2.44	(1.09)	1.00/.00	(.01)
Family instability	1.13	(1.15)	1.00/.00	(.01)
Maternal age at first child (years)	23.70	(5.80)	1.00/.00	(.00)
Less than high school diploma	11%	—	—	
High school diploma	25%	—	.95/–.05	(.03)
Some college or vocational training	32%	—	.97/–.03	(.03)
Bachelor's degree or higher	33%	—	.98/–.02	(.04)
Urban	71%	—	—	—
Suburban	12%	—	1.07/.06**	(.02)
Rural area	17%	—	1.04/.04+	(.02)
School and teacher characteristics			—	
Full-day kindergarten	75%	—	1.01/.01	(.02)
Public elementary school	89%	—	.94/–.06**	(.02)
Class size	19.79	(4.29)	1.00/.00	(.00)
Years teaching kindergarten	8.83	(7.96)	1.00/.00**	(.00)
Distance <1 mile	37%	—	—	—
Distance 1 to 2.5 miles	24%	—	.99/–.01	(.02)
Distance 2.6 to 5 miles	20%	—	1.00/.00	(.02)
Distance 5.1 to 10 miles	13%	—	1.00/.00	(.02)
Distance >10 miles	5%	—	.97/–.03	(.03)
Intercept	—	—	3.68/1.30	(.18)

Note. $n = 4{,}900$. Results aggregated across 20 complete imputed data sets. Outcome is not standardized. IRR = incidence rate ratio; *SE* = standard error; ECE = early childhood education.
$^{+}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

The third research aim assessed whether transition practices were associated with children's functioning in kindergarten. In each set of models, five separate OLS regression models were estimated to predict each of the five different child outcomes in kindergarten (prosocial skills, attention skills, child adjustment to school, reading skills, and mathematics skills). Each outcome was standardized for these models, so that coefficients indicate associations between a one-unit shift in the predictor and standard deviation (*SD*) unit shifts in the dependent variable. The first set of models included the index score of transition practices as the independent variable of interest. The second set of analyses included the seven individual practice indicator variables as the predictors of interest. Examinations of variance inflation factors (VIF) divulged no multicollinearity concerns for concurrent inclusion of all seven transition practices. VIF scores for the seven transition practices ranged from 1.03 to 1.16 across the child outcome models (mean VIF scores for all variables in these models = 1.60). The third and fourth sets of models incorporated interactions between centered measures of the transition practices and family income, first including the transition sum index, and second including the seven individual indicators. Because of high multicollinearity in models including interactions between individual transition indicators and family income (VIF scores ranged from 1.03 to 19.61), models were run including one interaction at a time.

All child outcomes models included the full set of child, family, and school covariates. Incorporation of such a rich set of covariates, many of which are associated with increased engagement in transition practices, helps to isolate unique associations between transition practices and children's successful adjustment to kindergarten. Each model also included the earlier measure of the child outcome variable, assessed the year prior to kindergarten, to control for the child's prior functioning. Because children's adjustment to kindergarten was not, by definition, assessed in preschool, parent report of children's shy and worried behaviors in the year prior to kindergarten was used as the lag in the model predicting children's adjustment. Inclusion of a lagged measure of child functioning adjusts for additional unmeasured factors that have a time-invariant effect on children's functioning (Cain, 1975), allowing us to interpret coefficients as effects of transition practices on changes in children's functioning over time (Kessler & Greenberg, 1981).

Results

Descriptive Results

Weighted descriptive data on the sample are presented in the first column of Table 1. Kindergarten teachers reported engaging in an average of just over three of seven transition practices. Frequencies of individual transition items demonstrate that transition practices reaching out to parents were the most common, with the vast majority of kindergarten teachers reporting phoning or sending information home to parents (90%) and having a parent orientation (84%). Similarly, the vast majority of teachers reported having children and parents visit the classroom (84%), but fewer had preschoolers spend time in the classroom (37%), and very few teachers engaged in home visits (4%) or structural practices, including staggered entry (14%) and shortened days for kindergarteners (10%).

Child, Family, Teacher, and School Characteristics Associated With School Transition Practices

The second column of Table 1 presents results from the Poisson regression model predicting the total transitions sum index. The incident rate ratios, or exponentiated coefficients, delineate the shift in estimated incident rates of transition practices for a one unit shift in the predictor. Few child and family characteristics were significantly associated with the number of transition practices. Greater time in kindergarten was associated with more transition practices, with a 1% increase in the number of practices with each additional month in kindergarten, suggesting that some of the practices occurred after the start of the school year. Children with a disability could expect to have kindergarten teachers who reported 5% more transition practices than the teachers of their typically developing peers. Native-born White children also had teachers reporting greater numbers of transition practices than their peers from African American, native-born Hispanic, immigrant Hispanic, and other immigrant families, with differences in incident rates ranging from 6% to 11%. The teachers of children residing in suburban communities also reported 7% greater numbers of transition practices than teachers of children in urban communities. Adjusting for these factors, family income and other measures associated with socioeconomic status and instability were not significantly associated with school transition practices. In terms of teacher and school characteristics, children attending public elementary schools had teachers reporting 6% fewer transition practices than their peers in private schools. In addition, greater teacher experience was associated with a negligible, but statistically significant, increase in transition practices.

Transition Practices Associated With Child Functioning

Children's behavioral functioning. Table 2 presents results from a series of OLS regression models testing associations between the transition practices index and children's functioning after the entry to kindergarten, adjusting for the full set of child, family, and school covariates and for children's functioning in preschool, assessed with the same individual measure used as the dependent variable (exception noted in analytic plan). Models included the kindergarten transition total sum index as the primary independent variable, and standardized measures of children's outcomes, such that coefficients indicate *SD* unit differences in child outcomes related to a one-unit difference in each predictor. Results show few significant links between transition practices and children's behavioral functioning in kindergarten. As one exception, the number of transition practices was associated with greater prosocial skills among children, with one additional transition practice associated with a small, but significant, .05 *SD* unit increase in children's prosocial skills. This finding holds even after adjusting for multiple comparisons using the Bonferroni correction (O. J. Dunn, 1961) to adjust the *p* value to 0.01.

The second set of models used the seven individual indicators of transition practices, with results presented Table 3. These models again controlled for the full set of child, family, and school

Table 2
OLS Models With the Transition Activities Index Predicting Child Outcomes

Variables	Prosocial skills		Attention skills		Positive adjustment		Reading		Mathematics	
	B	(*SE*)	*B*	(*SE*)	*B*	(*SE*)	*B*	(*SE*)	*B*	(*SE*)
Transition activities index	.05**	(.02)	.00	(.02)	.03	(.02)	.00	(.01)	.01	(.01)
Child and family characteristics										
Preschool wave prosocial skills	.38***	(.03)	—	—	—	—	—	—	—	—
Preschool wave attention skills	—	—	.47***	(.03)	—	—	—	—	—	—
Preschool wave positive adjustment	—	—	—	—	−.12***	(.03)	—	—	—	—
Preschool wave reading	—	—	—	—			.05***	(.00)	—	—
Preschool wave mathematics	—	—	—	—	—	—	—	—	.06***	(.00)
Child disability	−.18**	(.06)	−.19***	(.05)	−.09	(.06)	−.11**	(.04)	−.15***	(.04)
Male	−.27***	(.04)	−.27***	(.04)	−.20***	(.04)	−.02	.03	.05*	(.03)
Twin	.02	(.05)	.08	(.05)	.05	(.05)	−.01	.04	.02	(.04)
Low birth weight	−.03	(.05)	−.13**	(.05)	−.04	(.05)	−.07*	.03	−.13***	(.03)
In kindergarten in 2007	.04	(.07)	−.08	(.06)	.04	(.07)	.10*	(.05)	.06	(.04)
Age at kindergarten entry (months)	.01+	(.01)	.03***	(.01)	.01	(.01)	.01*	(.01)	.01**	(.00)
Months in kindergarten	.01	(.02)	−.01	(.01)	−.04+	(.02)	.17***	(.01)	.14***	(.01)
Siblings in same school	.05	(.04)	.11**	(.04)	.05	(.04)	.00	(.03)	.06*	(.03)
No EEC							—	—	—	—
Home ECE in child's home	−.06	(.11)	−.11	(.11)	−.01	(.13)	.01	(.08)	.05	(.07)
Home ECE in other home	.02	(.08)	.05	(.08)	.01	(.08)	−.04	(.06)	.02	(.05)
Center ECE in school	−.02	(.07)	−.1	(.06)	−.12+	(.07)	.05	(.05)	−.04	(.05)
Center ECE in other	.08	(.06)	−.05	(.06)	−.05	(.06)	.05	(.04)	.02	(.04)
Native White	—	—	—	—	—	—	—	—	—	—
Native Black	.08	(.07)	.06	(.06)	−.06	(.07)	−.02	(.04)	−.16+	(.04)
Native Hispanic	−.03	(.07)	−.05	(.07)	−.01	(.08)	.02	(.05)	−.08+	(.05)
Native American	−.13	(.13)	−.01	(.15)	−.26	(.19)	−.12+	(.07)	−.25	(.14)
Native multiple races	−.03	(.10)	−.03	(.11)	−.02	(.08)	−.03	(.07)	.01	(.06)
Asian	−.08	(.08)	.01	(.08)	.05	(.09)	.12+	(.06)	.06	(.06)
Non-Hispanic immigrant	−.09	(.09)	−.13	(.09)	.05	(.09)	.03	(.07)	−.04	(.06)
Hispanic immigrant	.09	(.10)	−.01	(.09)	−.12	(.10)	.01	(.07)	−.08	(.07)
Non-English household	.03	(.09)	.16+	(.08)	.12	(.09)	−.02	(.06)	−.07	(.06)
Family income ($10,000s)	.01	(.01)	.01	(.00)	.01*	(.01)	.00	(.00)	.00	(.00)
Maternal employment	−.03	(.04)	−.01	(.04)	.03	(.05)	.02	(.03)	−.04	(.03)
Maternal depression	−.15**	(.06)	−.15**	(.06)	−.13*	(.06)	.03	(.04)	.02	(.04)
Married parents	.15**	(.06)	.08	(.05)	−.08	(.06)	.09*	(.04)	.05	(.04)
Additional adults in household	−.04	(.05)	−.03	(.05)	−.05	(.05)	−.05	(.04)	−.03	(.03)
Children in household	−.04*	(.02)	−.03	(.02)	−.02	(.02)	−.03*	(.01)	.00	(.01)
Family instability	−.04*	(.02)	−.06**	(.02)	−.06**	(.02)	−.04**	(.01)	−.03**	(.01)
Maternal age at first child (years)	.00	(.00)	.01*	(.00)	−.01	(.00)	.00	(.00)	.00	(.00)
Less than high school diploma	—	—	—	—	—	—	—	—	—	—
High school diploma	−.02	(.07)	.02	(.07)	.05	(.08)	.16**	(.05)	.12*	(.05)
Some college or vocational training	.04	(.08)	.06	(.08)	.03	(.08)	.20***	(.06)	.17**	(.05)
Bachelor's degree or higher	.09	(.09)	.14	(.09)	−.06	(.10)	.31***	(.06)	.27***	(.06)
Urban	—	—	—	—	—	—	—	—	—	—
Suburban	.04	(.06)	.04	(.05)	−.07	(.07)	−.04	(.05)	−.03	(.04)
Rural area	.06	(.06)	−.01	(.06)	.06	(.06)	−.07	(.04)	−.04	(.04)
School and teacher characteristics										
Full-day kindergarten	.04	(.05)	.02	(.04)	−.14	(.05)	.12**	(.03)	.04	(.03)
Public elementary school	−.10	(.06)	.03	(.06)	.02	(.07)	.1**	(.04)	−.05	(.04)
Class size	.00	(.00)	.01	(.00)	.00	(.00)	.00	(.00)	.01*	(.00)
Years teaching kindergarten	.00	(.00)	.01**	(.00)	.00	(.00)	.00	(.00)	.00	(.00)
Distance <1 mile	—	—	—	—	—	—	—	—	—	—
Distance 1 to 2.5 miles	−.1*	(.05)	−.07	(.05)	.00	(.05)	−.01	(.04)	−.02	(.03)
Distance 2.6 to 5 miles	−.07	(.05)	−.04	(.05)	−.05	(.05)	.02	(.04)	−.01	(.04)
Distance 5.1 to 10 miles	−.07	(.06)	−.07	(.06)	.03	(.06)	−.02	(.05)	−.07	(.04)
Distance >10 miles	−.13	(.09)	−.08	(.09)	−.01	(.09)	−.07	(.06)	−.04	(.06)
Intercept	.15	(.13)	.10	(.12)	.29	(.13)	−.45	(.09)	−.11	(.09)
R^2 range	.163–.170		.248–.251		.047–.051		.547–.598		.604–.613	
R^2 average	.1652		.250		.049		.552		.610	
F-score range	11.08–12.22		24.12–24.92		2.76–3.01		77.15–80.31		91.11–96.29	
F-score average	11.46		24.47		2.87		78.23		94.27	

Note. $n = 4{,}900$. Results aggregated across 20 complete imputed data sets. All outcome variables are standardized. OLS = ordinary least squares; *SE* = standard error; ECE = early childhood education.
$^{+}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Table 3
OLS Models With Individual Transition Practices Predicting Child Outcomes

Variables	Prosocial skills		Attention skills		Positive adjustment		Reading		Mathematics	
	B	(*SE*)	*B*	(*SE*)	*B*	(*SE*)	*B*	(*SE*)	*B*	(*SE*)
Preschoolers visit classroom	.03	(.04)	.00	(.04)	.11*	(.04)	−.04	(.03)	−.01	(.03)
Parents and children visit class	.05	(.05)	−.06	(.05)	.01	(.06)	−.02	(.04)	.02	(.04)
Teacher home visits	−.04	(.13)	−.11	(.12)	.10	(.09)	−.04	(.09)	.02	(.07)
Send information to parents	.13*	(.06)	.02	(.06)	.02	(.07)	.06	(.05)	.00	(.04)
Parent orientation	−.03	(.05)	.02	(.05)	−.02	(.06)	.12**	(.04)	.12**	(.04)
Staggered entry in small groups	.02	(.06)	.02	(.06)	−.07	(.07)	−.03	(.04)	−.07+	(.04)
Shortened days	.10	(.06)	.04	(.06)	.02	(.07)	.00	(.05)	.08*	(.04)
Intercept	−.01	(.15)	.12	(.14)	.22	(.15)	−.55	(.11)	−.21	(.10)
R^2 range	.164–.171		.249–.253		.049–.054		.550–.561		.607–.616	
R^2 average	.166		.251		.052		.555		.613	
F-score range	9.78–10.79		21.52–22.31		2.67–2.87		69.13–72.42		81.71–87.80	
F-score average	10.16		21.84		2.75		70.64		85.23	

Note. $n = 4{,}900$. Models included the full set of child, family, school, and teacher covariates shown in Table 2, with coefficients for covariates not shown. Results aggregated across 20 complete imputed data sets. All outcome variables are standardized. OLS = ordinary least squares; *SE* = standard error.
$^{+} p < .10$. $^{*} p < .05$. $^{**} p < .01$. $^{***} p < .001$.

covariates and the child functioning lag. Again, few significant results emerged. Preschoolers spending time in the kindergarten class was associated with a .11 *SD* unit increase in children's school adjustment, and sending information home to parents was associated with a .13 *SD* unit increase in prosocial behaviors. However, when adjusting for multiple comparisons using the Bonferroni correction (O. J. Dunn, 1961), these results no longer reached statistical significance. Across both sets of models, school transitions were not significantly associated with children's attention skills in kindergarten.

Children's cognitive functioning. Table 2 and Table 3 also show results from the models predicting children's cognitive functioning after entry to kindergarten. In the first set of models in Table 2, no significant links emerged between the total transition practices index and children's reading or mathematics skills. However, consideration of individual transition practices, with models shown in Table 3, found that provision of parent orientations was associated with .12 *SD* unit increases in both children's reading skills and math skills. These findings are robust to the Bonferroni correction adjusting for multiple comparisons, suggesting that to enhance children's academic outcomes, practices focused on parents are more important that other individual practices, or simply more practices. In addition, shortened days for kindergarteners was associated with a .08 *SD* unit increase in math skills, yet this finding was not statistically significant when adjusting for multiple comparisons.

In the behavioral and cognitive models, significant covariates confirm the importance of accounting for other child, family, and school characteristics. Across all models, preschool functioning was strongly positively associated with children's kindergarten functioning, whereas child disability and low birth weight were often negatively associated. Family instability was negatively associated with all arenas of functioning. In addition, maternal depression showed significant negative associations with children's behavioral outcomes, and higher parent education was positively associated with children's cognitive outcomes.

Interactions between transition practices and family income. The final set of models tested whether associations between kindergarten transition practices and children's functioning were moderated by family income. We first tested interactions between family income and the total transitions sum index, and next tested interactions between income and each of the individual transition practices in separate models. After adjusting for multiple comparisons for five different outcomes, using the Bonferroni correction to get an adjusted *p* value of 0.01 (O. J. Dunn, 1961), no significant interactions emerged between the transition index and family income (results presented in the online supplemental materials). Similarly, no significant interactions emerged between family income and individual transition practice indicator variables, suggesting that links between transition practices and children's functioning were similar across the family income distribution.

Discussion

Successful transitions to kindergarten have both short-term and long-term implications for children and their families (Cowan et al., 2005; Entwisle & Alexander, 1993; Ladd et al., 2000; Ladd & Price, 1987; Pianta & Kraft-Sayre, 2003). As such, it is essential to understand how schools can help to ease this process for kindergarteners, supporting smooth and positive transitions. This study adds to the literature by exploring the transition practices used by kindergarten teachers; assessing the child, family, teacher, and school characteristics associated with greater use of transition practices; and delineating how engagement in different practices relates to children's social and academic adjustment in kindergarten in a nationally representative sample of children.

Prevalence and Correlates of Transition Practices

Through assessment of children who entered kindergarten in 2006–2007, we provide an updated picture of the transition practices reported by their kindergarten teachers. On average, we found that kindergarten teachers reported engaging in just over three of seven transition practices, with outreach to parents (i.e., sending information home and holding parent orientations) and children and parents visiting the classroom as the most common practices.

Structural changes to the school schedule in the beginning of the year such as staggered entry or shortened days to ease the transition were much less frequent. To gain a better understanding of the contexts in which teachers were engaging in greater transition practices, we assessed how a broad range of child, family, school, and teacher characteristics were associated with transition practices. Results showed that teachers with more experience, as well as those teaching in private schools, reported engaging in significantly more transition practices than their peers.

These models also found that children in African American, Hispanic, and immigrant families, as well as those living in urban areas, had teachers who reported fewer transition practices than those of their peers. This is consistent with past research that has shown that the use of transition practices differs across schools and the demographics of the children they serve (Early et al., 1999; Love et al., 1992; Pianta et al., 1999; Schulting et al., 2005). In contrast to prior research considering links between risk factors and transition practices (LoCasale-Crouch et al., 2008; Schulting et al., 2005), we did not find that family income was associated with the prevalence of transition practices, perhaps suggesting that bivariate connections between income and transitions may be driven by differences related to other family characteristics such as race and ethnicity, nativity, and urbanicity.

Transition Practices and Child Outcomes

This study also sought to extend the very limited prior research base linking kindergarten transition practices to successful kindergarten functioning among children. Expanding on past research, this study considered both cognitive and behavioral aspects of children's functioning in kindergarten, reported by parents, teachers, and direct assessments. Moreover, using longitudinal data, our models were able to adjust for a broad range of covariates as well as for earlier measures of children's functioning, central techniques to help isolate unique associations between transition practices and children's kindergarten functioning and to limit concerns over unmeasured heterogeneity or selection bias. Still, we caution that the work is correlational, and unmeasured factors may have continued to bias results.

Prospective models predicting children's functioning in kindergarten found limited support for the hypothesis that broader use of transition practices would support children's adjustment to kindergarten. Specifically, results found that use of more transition activities by kindergarten teachers was predictive of enhanced prosocial skills among children, but was not associated with children's attention skills, positive adjustment, or reading and math skills in kindergarten. These results provide limited evidence to support the "more is better" view of transition practices in which engaging in multiple different types of practices will provide the most support to parents and children, thereby enhancing positive child functioning in kindergarten.

We also assessed whether specific types of transition practices were linked to particular aspects of children's functioning. Here, results found that holding parent orientations was associated with increases in children's math and reading scores. These results suggest that sharing expectations with parents concerning the academic demands and practices of kindergarten might help to support children's academic skill gains. Parent orientations often relay expectations regarding daily reading at home, for example, and introduce parents to concepts such as homework folders and materials, which they are expected to review with their children nightly. Additional research is needed to further explore these potential mechanisms, seeking to elucidate the messages that are relayed in parent orientations, and to delineate how parents receive and interpret such information. Together, these results support a growing base of research supporting the importance of family–school connections in promoting school success among children (Pomerantz et al., 2007).

Although our results replicated and extended prior work showing the primary importance of parent orientations, results did not replicate prior research that has found transition practices to be more beneficial for children at risk because of limited family economic resources (LoCasale-Crouch et al., 2008; Schulting et al., 2005). In our models, children from lower income families neither had teachers reporting fewer transition practices, nor showed greater or lesser benefits from such practices than their more economically advantaged peers. One possible explanation for this inconsistency is simply that our analytic models more fully adjusted for a variety of other correlated characteristics of families and schools.

Limitations and Directions for Future Research

This study adds to the literature by providing an updated portrait of the transition practices used by kindergarten teachers and their associations with both social and academic functioning among new kindergarteners. Yet our understanding of the use and utility of transition practices remains somewhat limited. This and other past studies asked teachers about the general practices they engaged in for children in their class, but did not specify which children and parents actually participated in the activities. In addition, teachers reported only on whether or not they engaged in each transition practice, providing no information on the quality, frequency, or intensity of the activities. For example, when teachers reported that they did home visits, it is not clear whether home visits were conducted for all children in their class or just specific children, or the content and nature of the visit. Past research has also suggested that there are many barriers teachers face to engaging in transition practices, including lack of time and pay during the summer months and getting class lists too late (Early, Pianta, Taylor, & Cox, 2001). Future research should seek to more richly describe the reach and intensity of transition practices as well as optimal structures for supporting teachers in their full use.

In addition, this study looked primarily at horizontal transition practices that connect children and families to the elementary schools they are entering, and did not address the early education settings they were coming from or the transition practices that may have been led by children's early education settings. Following the developmental ecological transition to kindergarten model (Rimm-Kaufman & Pianta, 2000), it is important for more research to address connections made between the early childhood settings children are leaving and the elementary schools they are entering—so-called vertical transition practices. Updated research is needed to explore the barriers teachers face in engaging in both horizontal and vertical transition practices and their relationships to both short and long-term functioning among children.

Finally, it is essential to reiterate some limitations of the data and statistical methods used in this research. Although we assessed

a large sample of children, followed prospectively, and incorporated sampling weights which were designed to allow generalizability to all children born in the United States in 2001, it is possible that weights did not fully adjust for all nonresponse and attrition bias. Moreover, although we adjusted for a broad range of child, family, teacher, and school covariates, our models remain correlational, and cannot support causal conclusions. Moreover, it is important to note that the size of the significant effects unearthed in this research was consistently small.

Implications for Policy and Practice

Despite the aforementioned limitations, this study provides insight into the types of transition practices that may help support children's positive adjustment to school. In an environment in which school resources are limited, and districts, schools, and teachers must make choices about how to spend their financial and human resources effectively, educators are pushed to make strategic decisions about the practices they engage in to best support children's successful transitions to school. This study extends results from the very limited prior research on transition practices engaged in by kindergarten teachers in the United States (Ahtola et al., 2011; LoCasale-Crouch et al., 2008; Schulting et al., 2005) to suggest that a broader use of diverse transition practices (more practices) may support children's prosocial behaviors with peers. Our findings also suggest that outreach specifically to parents through parent orientations may be a key transition practice for supporting children's academic success in both early reading and mathematics. Moreover, the relationships between transitions practices and children's social and academic success in kindergarten were not moderated by income in a nationally representative sample, suggesting that transition practices are beneficial for children across the income spectrum.

References

Ahtola, A., Silinskas, G., Poikonen, P.-L., Kontoniemi, M., Niemi, P., & Nurmi, J.-E. (2011). Transition to formal schooling: Do transition practices matter for academic performance? *Early Childhood Research Quarterly, 26,* 295–302. http://dx.doi.org/10.1016/j.ecresq.2010.12.002

Cain, G. G. (1975). Regression and selection models to improve nonexperimental comparisons. In C. A. Bernett & A. A. Lumsdiane (Eds.), *Evaluation and experiment* (pp. 297–317). New York, NY: Academic Press. http://dx.doi.org/10.1016/B978-0-12-088850-4.50009-4

Chernoff, J. J., Flanagan, K. D., McPhee, C., & Park, J. (2007). *Preschool: First findings from the third follow-up of the Early Childhood Longitudinal Study, Birth cohort (ECLS-B).* (NCES 2008–025). Washington, DC: National Center for Education Statistics.

Coley, R. L., Votruba-Drzal, E., Collins, M., & Cook, K. D. (2016). Comparing public, private, and informal preschool programs in a national sample of low-income children. *Early Childhood Research Quarterly, 36,* 91–105. http://dx.doi.org/10.1016/j.ecresq.2015.11.002

Coley, R. L., Votruba-Drzal, E., Miller, P. L., & Koury, A. (2013). Timing, extent, and type of child care and children's behavioral functioning in kindergarten. *Developmental Psychology, 49,* 1859–1873. http://dx.doi.org/10.1037/a0031251

Cowan, P. A., Cowan, C. P., Ablow, J. C., Johnson, V. K., & Measelle, J. R. (2005). *The family context of parenting in children's adaptation to elementary school.* Mahwah, NJ: Erlbaum.

Cowan, P. A., & Heming, G. (2005). How children and parents fare during transition to school. In P. A. Cown, C. P. Cowan, J. C. Ablow, V. K. Johnson, & J. R. Measelle (Eds.), *The family context of parenting in children's adaptation to elementary school* (pp. 79–118). Mahwah, NJ: Lawrence Erlbaum Associates.

Dearing, E., Kreider, H., & Weiss, H. (2008). Increased family involvement in school predicts improved child-teacher relationships and feelings about school for low-income children. *Marriage & Family Review, 43,* 226–254. http://dx.doi.org/10.1080/01494920802072462

Duncan, S. E., & DeAvila, E. A. (1998). *PreLAS.* Monterey, CA: CTB/McGraw-Hill.

Dunn, L. M., & Dunn, L. M. (1997). *Examiner's manual for the PPVT-III: Peabody picture vocabulary test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56,* 52–64. http://dx.doi.org/10.1080/01621459.1961.10482090

Early, D., Pianta, R. C., & Cox, M. J. (1999). Kindergarten teachers and classrooms: A transition context. *Early Education and Development, 10,* 25–46. http://dx.doi.org/10.1207/s15566935eed1001_3

Early, D. M., Pianta, R. C., Taylor, L. C., & Cox, M. J. (2001). Transition practices: Findings from a national survey of kindergarten teachers. *Early Childhood Education Journal, 28,* 199–206. http://dx.doi.org/10.1023/A:1026503520593

Eckert, T. L., McIntyre, L. L., DiGennaro, F. D., Arbolino, L. A., Perry, L. J., & Begeny, J. C. (2008). Researching the transition to kindergarten for typically developing children: A literature review of current processes, practices and programs. In D. H. Molina (Ed.), *School psychology: 21st century issue and challenges* (pp. 235–252). Hauppauge, NY: Nova Sciences.

Entwisle, D. R., & Alexander, K. L. (1993). Entry into school: The beginning school transition and educational stratification in the United States. *Annual Review of Sociology, 19,* 401–423. http://dx.doi.org/10.1146/annurev.so.19.080193.002153

Erikson, E. H. (1950). *Child and society* (2nd ed.). New York, NY: W. W. Norton & Company, Inc.

Ginsburg, H. P., & Baroody, A. J. (2003). *TEMA-3: Test of early mathematics ability* (3rd ed.). Austin, TX: Pro-Ed.

Gresham, F. M., Elliott, S. N., & Black, F. L. (1987). Teacher-rated social skills of mainstreamed mildly handicapped and nonhandicapped children. *School Psychology Review, 16,* 78–88.

Henderson, A. T., & Berla, N. (Eds.). (1994). *A new generation of evidence: The family is critical to student achievement.* Washington, DC: National Committee for Citizens in Education.

Kagan, S. L., & Neuman, M. J. (1998). Lessons from three decades of transition research. *The Elementary School Journal, 98,* 365–379. http://dx.doi.org/10.1086/461902

Kessler, R., & Greenberg, E. F. (1981). *Linear panel analysis: Models of quantitative change.* New York, NY: Academic Press.

Kull, M., Coley, R. L., & Lynch, A. D. (2015). The roles of instability and housing in low-income families' residential mobility. *Journal of Family and Economic Issues.* Advance online publication. http://dx.doi.org/10.1007/s10834-015-9465-0

Ladd, G. W., Buhs, E. S., & Seid, M. (2000). Children's initial sentiments about kindergarten: Is school liking an antecedent of early classroom participation and achievement? *Merrill-Palmer Quarterly, 46,* 255–279.

Ladd, G. W., & Price, J. M. (1987). Predicting children's social and school adjustment following the transition from preschool to kindergarten. *Child Development, 58,* 1168–1189. http://dx.doi.org/10.2307/1130613

La Paro, K. M., Kraft-Sayre, M., & Pianta, R. C. (2003). Preschool to kindergarten transition activities: Involvement and satisfaction of families and teachers. *Journal of Research in Childhood Education, 17,* 147–158. http://dx.doi.org/10.1080/02568540309595006

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83,* 1198–1202. http://dx.doi.org/10.1080/01621459.1988.10478722

LoCasale-Crouch, J., Mashburn, A. J., Downer, J. T., & Pianta, R. C. (2008). Pre-kindergarten teachers' use of transition practices and children's adjustment to kindergarten. *Early Childhood Research Quarterly, 23,* 124–139. http://dx.doi.org/10.1016/j.ecresq.2007.06.001

Lonigan, C., Wagner, R., Torgeson, J., & Rashotte, C. (2002). *Preschool Comprehensive Test of Phonological & Print Processing* [Unpublished test]. Department of Psychology, Florida State University, Tallahassee, FL.

Love, J. M., Logue, M. E., Trudeau, J. V., & Thayer, K. (1992). *Transitions to kindergarten in American schools.* Portsmouth, NH: U.S. Department of Education.

Merrell, K. W. (2003). *Preschool and kindergarten behavior scales* (2nd ed.). Austin, TX: PRO-ED.

National Education Goals Panel. (1998). *Ready schools.* Washington, DC: U.S. Government Printing Office.

Pianta, R. C., Cox, M. J., & Snow, K. L. (2007). *School readiness and the transition to kindergarten in the era of accountability.* Baltimore, MD: Brookes.

Pianta, R. C., Cox, M. J., Taylor, L., & Early, D. (1999). Kindergarten teachers' practices related to transition into schools: Results of a national survey. *The Elementary School Journal, 100,* 71–86. http://dx.doi.org/10.1086/461944

Pianta, R. C., & Kraft-Sayre, M. (2003). *Successful kindergarten transition: Your guide to connecting children, families, and schools.* Baltimore, MD: Brookes.

Pianta, R. C., & Walsh, D. J. (1996). *High-risk children in schools: Creating sustaining relationships.* New York, NY: Routledge.

Pomerantz, E. M., Moorman, E. A., & Litwack, S. D. (2007). The how, whom, and why of parents' involvement in children's academic lives: More is not always better. *Review of Educational Research, 77,* 373–410. http://dx.doi.org/10.3102/003465430305567

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385–401. http://dx.doi.org/10.1177/014662167700100306

Ramey, C. T., & Ramey, S. L. (2010). The transition to school: Concepts, practices, and needed Research. In S. L. Kagan & K. Tarrant (Eds.), *Transitions for young children: Creating connections across early childhood systems* (pp. 33–44). Baltimore, MD: Brookes.

Ramey, C. T., Ramey, S. L., & Lanzi, R. G. (2006). Children's health and education. In I. Siegel & A. Renninger (Eds.), *The handbook of child psychology* (Vol. 4, pp. 864–892). Hoboken, NJ: Wiley.

Reynolds, A. J. (2004). Research on early childhood interventions in the confirmatory mode. *Children and Youth Services Review, 26,* 15–38. http://dx.doi.org/10.1016/j.childyouth.2003.11.001

Rimm-Kaufman, S. E., & Pianta, R. C. (2000). An ecological perspective on the transition to kindergarten: A theoretical framework to guide empirical research. *Journal of Applied Developmental Psychology, 21,* 491–511. http://dx.doi.org/10.1016/S0193-3973(00)00051-4

Rimm-Kaufman, S. E., Pianta, R. C., & Cox, M. J. (2000). Teachers' judgments of problems in the transition to kindergarten. *Early Childhood Research Quarterly, 15,* 147–166.

Royston, P. (2005). Multiple imputation of missing values: Update of ice. *The Stata Journal, 5,* 527–536.

Schulting, A. B., Malone, P. S., & Dodge, K. A. (2005). The effect of school-based kindergarten transition policies and practices on child academic outcomes. *Developmental Psychology, 41,* 860–871. http://dx.doi.org/10.1037/0012-1649.41.6.860

Snow, K. L. (2006). Measuring school readiness: Conceptual and practical considerations. *Early Education and Development, 17,* 7–41. http://dx.doi.org/10.1207/s15566935eed1701_2

Snow, K., Derecho, A., Wheeless, S., Lennon, J., Rosen, J., Rogers, J., . . . Einaudi, P. (2009). *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), kindergarten 2006 and 2007 data file user's manual (2010–010).* Washington, DC: U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences.

Vernon-Feagans, L., Garrett-Peters, P., Willoughby, M., Mills-Koonce, R., & the Family Life Project Key Investigators. (2012). Chaos, poverty, and parenting: Predictors of early language development. *Early Childhood Research Quarterly, 27,* 339–351. http://dx.doi.org/10.1016/j.ecresq.2011.11.001

Received May 26, 2015
Revision received May 2, 2016
Accepted May 6, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 2, 178–190

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000118

Individual and Class Norms Differentially Predict Proactive and Reactive Aggression: A Functional Analysis

Karin S. Frey
University of Washington

Zoe Higheagle Strong
Washington State University

Adaurennaya C. Onyewuenyi
University of Washington

Theory and research using a social-information processing framework indicate that reward-focused (proactive) aggression has different social consequences than defense-focused (reactive) aggression. Students use norms that identify expected and socially approved behaviors as guides to their own actions. Differences in social-cognitive processing characteristics and social status linked to each type of aggression may increase the relevance of some normative sources relative to others. This study fills a gap in the literature by examining the contributions of personal beliefs, classroom beliefs, and classroom rates of aggression to future proactive and reactive aggression. During fall and spring, we observed students' aggression on school playgrounds using a random subsample (n = 254) of consented students from 35 classrooms (Grades 3–6). We calculated classroom rates of proactive and reactive aggression from fall observations. Classroom means for beliefs endorsing retaliation were calculated from surveys of 536 students. Results of multilevel analyses revealed, as hypothesized, that personal beliefs predicted high rates of students' proactive aggression, but not reactive aggression. Classroom beliefs predicted high rates of students' reactive but not proactive aggression. Students in classrooms with high rates of fall proactive aggression showed low spring rates of both types of aggression. In contrast, students in classrooms with high rates of fall reactive aggression displayed high spring rates of proactive and reactive aggression. The latter pattern may represent classrooms in which students continue to struggle against status inequities. The discussion examines how inequities may impact intervention efforts.

Keywords: proactive aggression, reactive aggression, retaliatory beliefs, classroom norms, peer influence

Aggression is a chronic problem in our nation's schools, leading many educators to adopt aggression reduction programs. These programs often aim to promote normative beliefs and expectations among students that aggression is unacceptable or ineffective. Effectiveness varies, however (Ansary, Elias, Greene, & Green, 2015; Ttofi & Farrington, 2011), and efficacy of intervention practices may depend on the functions served by specific normative beliefs and aggressive behaviors. Functional distinctions between proactive aggression (also called instrumental or reward-focused) and reactive aggression (also called retaliatory or defensive) are associated with different patterns of social goals, outcome expectations, and self-regulatory abilities (see reviews and meta-analyses; Card & Little, 2006; Hubbard, McAuliffe, Morrow, & Romano, 2010; Polman, de Castro, Koops, van Boxtel, & Merk, 2007). Thus, normative beliefs at the personal and classroom levels and typical patterns of classroom behavior may differentially predict changes in the two types of aggression.

Despite a rich history of theoretical and empirical work on normative influences (e.g., Bandura, Caprara, Barbaranelli, Pastorelli, & Regalia, 2001; Huesmann, 1988; Henry et al., 2013) and on functional distinctions in aggression (Crick & Dodge, 1994; Dodge & Coie, 1987; Marsee et al., 2014), few studies have examined the specific contributions of aggressive norms to proactive and reactive aggression. Key variations in the social-cognitive processes and the social context associated with each aggression type suggest that the relative influence of social norms will differ in systematic ways. This study employed Dodge's social information processing model (SIP; Fontaine & Dodge, 2006) as a framework for examining the contributions of specific social norms to proactive and reactive aggression.

This article was published Online First June 2, 2016.

Karin S. Frey, Educational Psychology, College of Education, University of Washington; Zoe Higheagle Strong, College of Education, Washington State University; Adaurennaya C. Onyewuenyi, Educational Psychology, College of Education, University of Washington.

We thank Leihua V. Edstrom, Miriam K. Hirschstein, and Jennie Snell for their invaluable contributions to the original evaluation study, to Committee for Children for funding data collection, and to Debra Pepler and Wendy Craig for their sage advice. Thanks also to Tom Dorrance, Leihua Edstrom and our review team for helpful critiques of the manuscript.

Correspondence concerning this article should be addressed to Karin S. Frey, Educational Psychology, College of Education, Miller Hall, University of Washington, Box 353600, Seattle, WA 98195-3600. E-mail: karinf@uw.edu

Theory and Research on Proactive and Reactive Aggression

SIP links the selection and cnactment of social behaviors to six elements in a nearly instantaneous decision-making process (Fontaine & Dodge, 2006). Each step in the process can be performed with varying degrees of adequacy, depending on situational and personal characteristics. In Step 1, a student notices social cues (e.g., a derisive tone of voice, the presence of bystanders), which form the basis for social inferences in Step 2 (e.g., "This person wants to embarrass me publicly."). In Step 3, a student identifies personal goals (e.g., reduce negative emotions, reclaim lost status). In Step 4, a student generates alternative ways to achieve a goal (e.g., run away, hit the kid, wait and retaliate anonymously). Students evaluate alternatives in Step 5 based on beliefs that reflect sociomoral values, self-efficacy (e.g., "I'm not a good fighter."), and expectations of success in a specific situation (e.g., "I'm alone, but that person's friends are here"). Finally, in Step 6, students perform the selected action. These steps are performed in emotional contexts that vary in valence and intensity.

A high level of emotional arousal is likely to interfere with decision-making processes. This leads to actions that may be based on inadequate attention to social cues, snap judgments about others' intentions, and failure to evaluate the adequacy of possible responses. Highly emotional, impulsive responses are typical of reactive aggression (Card & Little, 2006; Hubbard et al., 2010; Polman et al., 2007). They may be provoked by unjust or otherwise aversive situations. Moreover, reactive aggression is associated with biases toward encoding and interpreting social cues as hostile, and responses that are retaliatory.

Proactive aggression refers to intentional acts of aggression that are aimed at achieving a desired goal and are contingent on evaluation of consequences (Fontaine & Dodge, 2006). Proactive aggression is uniquely related to bullying: the repeated targeting of a person of lesser power (Fossati et al., 2009). Although proactive aggression may be accompanied by anger, emotion expression is likely to be regulated in ways that enable strategic decision-making and goal acquisition (Hubbard et al., 2002). Retaliation, for example, may be delayed in order to assure a successful outcome and an audience for the antagonist's humiliation (Dodge, 1991). Manipulation of peers (Little, Henrich, Jones, & Hawley, 2003) may also disguise the avenger's identity or incite allies to retaliate (Frey, Pearson, & Cohen, 2015; Garandeau & Cillessen, 2006; Xie, Swift, Cairns, & Cairns, 2002).

In general, proactive aggression appears to be associated with successful goal achievement more than reactive aggression. Reactive aggression is related to peer rejection (Evans, Fite, Hendrickson, Rubens, & Mages, 2015) and victimization (Salmivalli & Helteenvuori, 2007), while proactive aggression is associated with a dominant position in the social hierarchy (Pellegrini, Bartini, & Brooks, 1999; Pellegrini et al., 2011; Sijtsema, Veenstra, Lindenberg, Salmivalli, 2009; cf., Polman, de Castro, Thomaes, & van Aken, 2009). The status that often accompanies proactive aggression may be a key element of success as students navigate, influence and even co-opt classroom norms for behavior.

Normative Beliefs and Behavior

Normative influences include internal, personal beliefs, and those that are shared within groups such as classrooms. Personal beliefs become visible to classmates via gossip, advice, and exhortations. When particular values and beliefs are widely shared, they become part of the classroom culture (Hawley & Williford, 2015).

Normative *beliefs* are *injunctive* norms—what most people think *should* happen (Henry, 2008). They describe beliefs about moral and conventional obligations and constraints. Normative beliefs provide support for self-regulatory efforts as people try to act in ways congruent with personal and cultural definitions of a good person (Bandura et al., 2001). Normative *behavior,* on the contrary, is a *descriptive* norm, based on patterns of actions and reactions that typify social groups. Classmates can observe behaviors, such as aggressive acts that are common in specific situations, and the attendant rewards and punishments.

Common classroom patterns associated with aggression may acquire a veneer of perceived legitimacy as in "Everybody does it." (Ang, Ong, Lim, & Lim, 2010; Chang, 2004), even if they are not sanctioned by shared classroom beliefs. The responses of victims and bystanders to aggression contribute to classroom norms and outcome expectations by determining whether an aggressive act is viewed as efficacious, rewarding, or costly (Pellegrini, 2008). In order to better understand normative influences on aggression, we need to consider descriptive norms such as classroom rates of aggression, as well as injunctive norms such as beliefs regarding aggression at the personal and classroom level.

Past Research on Normative Contributions to Aggression

Previous research identified links between personal beliefs endorsing aggression and aggressive behavior in concurrent analyses (Kikas, Peets, Tropp, & Hinn, 2009; Werner & Nixon, 2005) and longitudinal analyses (Henry et al., 2000; Huesmann & Guerra, 1997; Werner & Hill, 2010). Physical and relational aggression generally show similar relationships to beliefs. Further, beliefs that justify bullying are linked to high levels of bullying behavior (Perren, Gutzwiller-Helfenfinger, Malti, & Hymel, 2012; Sentse, Veenstra, Kiuru, & Salmivalli, 2015).

Normative beliefs that are widely shared specify who is rewarded by aggression, how, and under what circumstances. Past research shows inconsistent links between classroom beliefs (usually calculated as the group mean of personal beliefs) and later aggressive behavior (Henry et al., 2000; Sentse et al., 2015; Werner & Hill, 2010).

The level of aggressive actions within schools, classrooms, and peer groups has also been shown to contribute to later student aggression (Mercer, McMillen, & DeRosier, 2009; Sentse et al., 2015; Thomas, Bierman, & Powers, 2011). Most multilevel studies have not included both normative beliefs and behaviors as predictors of later aggression. Those that have are inconclusive regarding the relative magnitude or possible causal ordering of each type of normative influence (e.g., Henry et al., 2000; Huesmann & Guerra, 1997; Sentse et al., 2015).

Norms Are Pertinent to Response Evaluation

According to the SIP model, normative influences derived from personal and classroom beliefs and classroom rates of aggression are most likely contribute to social decision-making during Step 5: response evaluation. Relatively stable mental structures such as beliefs about sociomoral acceptability, self-identity, and expectations of positive and negative outcomes form the basis for response evaluation (Fontaine & Dodge, 2006). Given the processing differences involved in proactive and reactive aggression, the relevance of norms in decision-making is likely to vary by the type of aggression and the type and source of the normative influence.

Normative Beliefs Endorsing Retaliation

This study focused specifically on personal and classroom beliefs regarding retaliation, which vary according to culture and geographic region (Cohen, 2001). It might be assumed that beliefs endorsing retaliation are pertinent only for reactive (i.e., retaliatory) aggression, but two lines of inquiry suggest otherwise. First, personal beliefs endorsing retaliation appear to disengage moral constructs against harming others (Bandura et al., 2001) with either proactive or reactive aggression. Endorsement of retaliation and justifications based on prior offense are associated with high levels of bullying, for example (Bradshaw, O'Brennan, & Sawyer, 2008; Perren et al., 2012). Further, group norms regarding retaliation are likely to affect the costs and rewards associated with aggression, as well as being reflected in the decision-making process. Thus, beliefs endorsing retaliation may be relevant for both proactive and reactive aggression. But as detailed next, contribution may vary with the source (personal or classroom beliefs), and a student's online processing adequacy when aggression becomes a possibility.

Do Norms Differentially Predict Students' Proactive and Reactive Aggression?

Despite increasing interest in ecological influences on aggression (see review, Pellegrini, 2008), the normative influences on students' proactive and reactive aggression have virtually been ignored. To our knowledge, only one study on young adults has examined the relationships of reactive and proactive aggression to personal beliefs (Bailey & Ostrov, 2008). Since it is unclear whether research on young adults can be applied to elementary classrooms, more investigation is warranted on normative influences, particularly in the school setting.

Elementary schoolchildren spend much of the school day with classmates, who are likely to be a potent source of normative influence. Whether their influence makes similar contributions to proactive and reactive aggression is a question of both theoretical and practical importance. There is a particular need for longitudinal data, since normative influences may significantly moderate intervention efficacy. Our study responds to this need by investigating the complex relationships between specific normative influences at the personal- and classroom-levels and subsequent aggressive actions. Consistent with behavioral ecological analyses (Pellegrini, 2008), we assume that classroom beliefs and classroom rates of proactive and reactive aggression will create costs and benefits that are differentially reflected in students' rates of proactive and reactive aggression.

Current Study

The data for the current study were collected as part of an evaluation of an antibullying program. We conducted fall and spring playground observations and student surveys of beliefs endorsing retaliation for students in Grades 3–6. These grades precede declines in the effectiveness of bullying prevention programs (Yeager, Fong, Lee, & Espelage, 2015), but coincide with developmental changes that are eventually likely to be relevant to program effectiveness. For example, approval for retaliation increases during these grades (Frey, Hirschstein, Edstrom, & Snell, 2009; Huesmann & Guerra, 1997), as does sensitivity to social cues (Blakemore & Mills, 2014).

To our knowledge, this is the first multilevel longitudinal study to distinguish between the contributions that normative beliefs and behaviors make to each type of aggression. Our hypotheses predict the separate contributions of personal beliefs, classroom beliefs, classroom rates of proactive aggression, and classroom rates of reactive aggression to students' proactive and reactive aggression. We next provide the rationale for each hypothesis. Hypotheses are also summarized in Figure 1.

Hypotheses for Proactive Aggression Reflect Goal Focus and Peer Dominance

Personal Beliefs Considered During Response Evaluation

The goal of proactive aggression is to obtain rewards with minimal cost. This strategic quality suggests that response evaluation, Step 5, will play an important role in the social decision-making process (Fontaine & Dodge, 2006). Stable mental structures such as personal beliefs will inform response evaluation and selection. For example, aggressive self-efficacy, a characteristic of proactive aggressors (Hubbard et al., 2010), predicts retaliation (Erdley & Asher, 1996). To someone with a goal of being socially dominant, provocations that go unpunished may imperil goal achievement and thus justify aggression (Boyd, 2014). Low impulsivity also means that aggressors who are predominantly proactive can maintain a revenge goal over time and strategize accordingly. With these considerations in mind, we predicted that personal beliefs endorsing retaliation would contribute to high rates of students' proactive aggression in the spring after accounting for prior aggression rates.

Leadership May Negate Contribution of Classroom Beliefs

As a well-regulated goal-directed behavior, proactive aggression may be relatively unaffected by in-the-moment peer pressure to retaliate (Thomas & McGloin, 2013). Successful proactive aggressors, however, will be cognizant of threats to their dominance goals if they lose peer support for their aggression (Veenstra, Lindenberg, Munniksma, & Dijkstra, 2010). Importantly, proactive aggressors try to shape group norms to their own specifications (Allen, Porter, & McFarland, 2006). Manipulating classmates' interpretation of events (Little et al., 2003) by appealing to retaliation norms may be an efficacious strategy. In speaking with allies, for example, proactive aggressors may recast their own aggression as a justifiable response to victim retaliation (Boyd, 2014). If proactive aggressors have a disproportionate influence on classroom beliefs, then

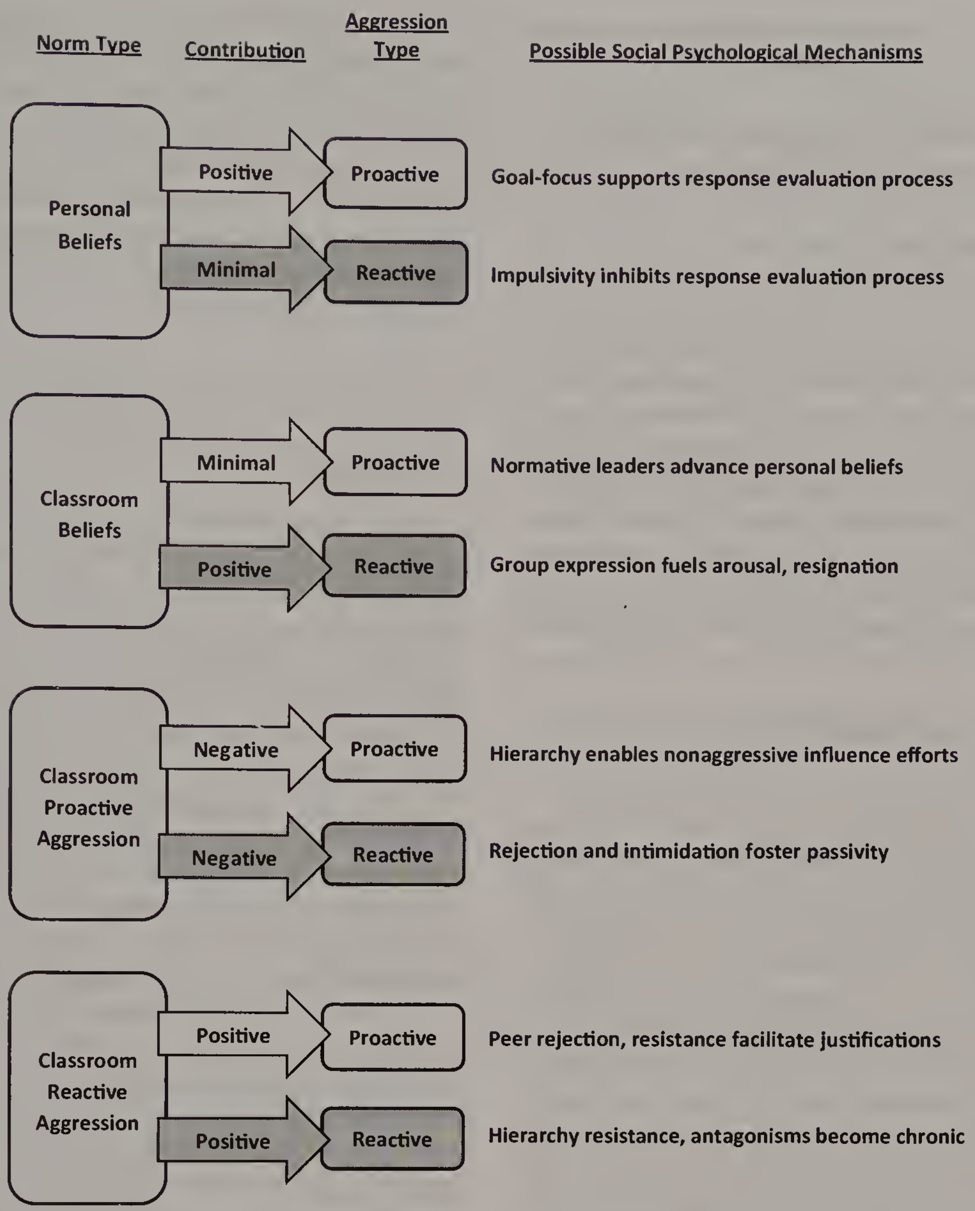


Figure 1. Model illustrates the hypothesized contributions of normative beliefs and classroom rates of aggression to subsequent personal rates of proactive and reactive aggression.

those beliefs are unlikely to contribute significantly to later proactive aggression beyond the contribution of personal beliefs (Sentse et al., 2015). Thus, we did not predict a significant relationship between classroom beliefs endorsing retaliation in the fall and students' proactive aggression in the spring.

Classroom Proactive Aggression and the Advantage of Power

Two lines of reasoning offer divergent hypotheses. On one hand, the elevated status enjoyed by proactive aggressors (Pellegrini et al., 2011) is a compelling reward. Past success derived from aggressive strategies may encourage continued high rates of proactive aggression (Sentse et al., 2015), particularly if competition for dominance continues (Garandeau & Cillessen, 2006). On the other hand, the establishment of a favorable dominance hierarchy at the beginning of the year may enable some proactive aggressors to desist (Faris & Felmlee, 2011). Mindful of peer support (Veenstra et al., 2010), they may prefer to dispense selective kindness and foster strategic alliances (Hawley, 2003; Pellegrini et al., 2011; Roseth et al., 2011). Such a shift in direction is easier to accomplish if proactive aggression confers protection against later aggression by classmates (Frey, Newman, & Onyewuenyi, 2014). These considerations suggest an alternative hypothesis—that high classroom rates of proactive aggression in the fall will contribute to low rates of students' proactive aggression in the spring.

Classroom Reactive Aggression May Enable Opportunistic Justifications

Reactive aggressors may be socially inept, enabling proactive aggressors to manipulate peer opinion (Garanadeau & Cillessen, 2006; Little et al., 2003) and the social rejection of reactive aggressors (Evans et al., 2015). Peers' lack of sympathy for reactive aggressors enable proactive aggressors to target them without fear of losing peer support (Veenstra et al., 2010). Thus, high levels of classroom reactive aggression may actually help proactive aggressors bully with impunity. Another consideration is that resistance to domination by proactive aggressors may take the form of reactive aggression. If reactive aggression rates are indicative of widespread resistance, proactive aggressors may deem it necessary to frequently display dominance aggressively. Both lines of reasoning suggest

that high classroom rates of reactive aggression in the fall will contribute to high rates of students' proactive aggression in the spring.

Hypotheses for Reactive Aggression Reflect Impulsivity and Peer Rejection

Impulsivity Inhibits Evaluation Based on Personal Beliefs

Although reactive aggressors may hold the same beliefs about retaliation as their peers, they are less likely to review such mental structures when deciding whether or not to retaliate. Response evaluation (Step 5) is often cursory or omitted entirely as a result of emotional dysregulation. Indeed, Arsenio and Gold (2006) find that reactively aggressive youth display normal moral development, but impulsivity restricts their ability to regulate their behavior on the basis of normative standards. Thus, we predicted that personal beliefs will not contribute to students' reactive aggression in the spring.

Expressed Classroom Beliefs May Fuel Arousal and Resignation

In contrast, impulsive youth appear to be particularly influenced by social norms (Thomas & McGloin, 2013), perhaps due to unmet needs for social acceptance (Evans et al., 2015). These youth, however, do not necessarily evaluate responses systematically in light of classroom norms. Classroom beliefs and expectations may be revealed in the exhortations of bystanders (e.g., when they are eager to see a fight). At such times, adolescents report emotional flooding and confusion (Farrell, Henry, Schoeny, Bettencourt, & Tolan, 2010), lessening constraints that might be provided by personal norms or even fear of the consequences. Furthermore, targets of aggression may feel that the classroom culture will eventually force retaliation, even if they do not view it as a desirable response or an effective deterrent (Farrell et al., 2010; de Castro, Verhulp, & Runions, 2012). Thus, we predicted that classroom beliefs endorsing retaliation will contribute to increased rates of students' reactive aggression later in the spring, even after accounting for prior aggression and personal beliefs.

Classroom Proactive Aggression and Peer Rejection May Foster Passivity

Domination by others is likely to stimulate high rates of aggressive reactivity, at least in the short-term. Over the school year, however, successful intimidation and domination may result in fear and despair on the part of frequent targets. Resistance may appear futile (Craig, Pepler, & Blais, 2007) even without careful evaluation of costs and benefits in any one instance. Over time, resignation and fear may reduce retaliation against more effective aggressors (Brendgen et al., 2013; Camodeca & Goossens, 2005). Therefore, we expected that high classroom rates of proactive aggression would predict low rates of students' reactive aggression in the spring.

Classroom Reactive Aggression May Become Chronic

Impulsivity makes it difficult for reactively aggressive children to evaluate and avoid rough play and arguments (Frey et al., 2014), activities that frequently escalate to aggression. High fall rates of reactive aggression may be indicative of classrooms with high conflict and mutual antagonisms. High rates of fall reactive aggression may also indicate classrooms in which resistance to domination has become part of the classroom climate (Hawley & Williford, 2015). In consideration of both possibilities, we predicted that high classroom rates of reactive aggression in the fall would contribute to high rates of students' reactive aggression in the spring.

Method

Participants

This study examined two waves of data collected during a randomized controlled trial that examined the effects of an anti-bullying program in two midsized cities in the United States Pacific northwest (Frey et al., 2009). Because the intervention was expected to change the relationship between variables, we selected only 35 classrooms in the three control schools for the study. School selection was based on educators' willingness to be randomly assigned to either an intervention or delayed control group. The percentage of children receiving free and reduced-price lunch at the schools ranged from 21% to 60%. Adherence to ethical guidelines was strictly enforced. Active parental consent was obtained for 64% of all students, and written child assent was obtained from students in Grades 4–6. A total of 536 students were surveyed regarding endorsement of retaliation in the fall and spring. Fall surveys from this sample were used to calculate means for classroom beliefs.

Ten children (usually five girls and five boys) in each classroom were randomly selected for playground observation. Of an initial total of 284 in the observation sample, 30 children left their schools during the study (10.6%) and 5 (1.8%) were excluded for incomplete observations (less than 35 min at both Time 1 and Time 2). *T*-tests comparing fall rates of proactive and reactive aggression in the retained sample with those lost to attrition indicated no significant differences. The final observed sample ($n = 254$; 122 girls) was 10.2% African American, 9.4% Asian American, 68.9% European American, 9.4% Hispanic American, and 2.0% Indigenous American. Students who spoke English as a second language comprised 11.0% of the sample, although observed conversations were exclusively in English. Students remained in the same classroom every day.

Data Collection Procedures

Endorsement of retaliation. The Endorsement of Retaliation Survey was used to measure personal and classroom beliefs. It is a group-administered four-point Likert scale (1 = *Do not agree*; 4 = *Agree a lot*). An initial array of 18 items included ones that examined endorsement of indirect and relational aggression (Crick & Grotpeter, 1995), as well as direct physical and verbal aggression (Huesmann & Guerra, 1997). After pretesting with 126 students, eight items with strong factor loadings (four each for direct and indirect retaliation) were administered to 309 students in Grades 3–6. This yielded factor loadings ranging from .50--.60, except for one item with a loading of .23. That item was dropped, forming a scale with four items specifying direct retaliation and

three specifying indirect retaliation. Examples of final items include, "It's okay to hit someone who hits you first," "It's okay to say something mean to someone who's pushing you around," "It's okay to say something mean about someone who says something mean about you," and "It's okay to stop talking to someone to get even." The scale showed moderate 10-week stability, $r(289) = .36$, $p < 01$. Internal consistency in the current study was .87.

Observation overview. Youth were observed on playgrounds from late September through December, and again from late March through June. Morning recess was 15 min long, and afternoon recess was typically 25 min long, enabling five focal child observation periods. Typically, each student was observed once a week for 5 min over 10 to 12 weeks, with the order of observation determined randomly each week. This schedule provided enough repetitions ($M = 10$, range 7–12) to reduce censoring effects (Stoolmiller, Eddy, & Reid, 2000) and staggered sampling times over the broadest range of conditions (e.g., preholiday, rainy weather). When conversations required extended proximity, observers reduced reactivity by periodically shifting positions and continuously keeping the children in sight with direct or peripheral vision. Children were minimally reactive, commenting that the observers "don't do anything."

The coding manual was used in conjunction with custom-programmed handheld devices. Coding in real time (Sackett, 1977), observers opened multiple screens in a response-contingent order, thereby reducing operator error. Screen 1 identified the actor. Screen 2 identified aggressive, nonaggressive, or bystander behaviors. Indicating aggression on Screen 2 automatically led to Screen 3 for coding aggressive function.

Observer training. Of 15 paid observers who were blind to the specific purpose of the study, 14 successfully completed training. Training Phase 1 (200 hr) covered ethical guidelines, operational definitions, borderline decision-making, error correction, and data collection protocols. Trainees coded videotaped playground behaviors and received immediate feedback. In order to advance to in vivo coding, each coder had to agree with a master coder with an overall minimum mean kappa of .70 on videotapes. Training Phase 2 (40–50 hr of playground coding) allowed time for children to habituate to the presence of observers. For at least eight hours, each observer coded simultaneously with a master coder, receiving feedback, and reviewing discrepancies. Prior to spring data collection, coders underwent 20 hr of booster training before advancing to Training Phase 2. In fall and spring, data collection started after kappas averaged at least .70. To prevent decay, master coders performed agreement checks (15% of sessions, $n = 868$) on coder accuracy throughout data collection.

Observer accuracy. Two coders were said to agree on a behavior when they indicated the same code within one second of each other. Agreement of qualifier codes for aggressive function was event- rather than time-dependent. Consistent with the exacting training, percentage agreements were above 89% and overall kappa was excellent ($\kappa = .80$). Kappas were also excellent when computed for separate behaviors (reported below) despite typically low levels found when infrequent events diverge from 0.5 (Xu & Lorber, 2014).

Aggression ($\kappa = .76$) was coded when a focal child hurt a peer with physical acts, threats, exclusion, or demeaning comments and gestures. Aggression included both face-to-face encounters (e.g., punching; saying, "You can't sit here.") and actions out of the target's direct awareness (e.g., eye-rolling behind a target's back, derogatory gossip, plotting to exclude or otherwise harm a person). Both conversation content and nonverbal cues (e.g., *significant looks* in the direction of the target) were used to identify derogatory gossip. Verbal statements could be coded as aggressive based on derogatory content even if the speaker was smiling or laughing. Similarly, coders distinguished between aggression and rough play. The latter was accompanied by mutual *felt* smiles and laughter. Bouts of rough play sometimes devolved into aggression, accompanied by shifts from positive to negative expressions on the part of at least one participant.

Aggressive function ($\kappa = .80$) was coded as *proactive* if initiated without any provocation apparent during the 5-min coding period. Aggression was coded as *reactive* if it appeared to be an impulsive retaliatory response, occurred during a disagreement (e.g., dispute about whether a player was *safe*), or immediately followed other types of provocation such as cutting in line.

Computation of Classroom Beliefs and Classroom Rates of Aggression

Although the study predicts the rates of students' aggression in the spring with a minimum of 35 min of observation time in both spring and fall, scores from the entire classroom were used to calculate means for classroom beliefs. Mean classroom beliefs were based on a minimum of seven cases ($M = 16.21$ cases, total $N = 584$).

In multilevel analyses, tests of fixed parameter coefficients are relatively robust when there is a small average number of cases per classroom (e.g., 4 to 5). Power to detect interactions of classroom- and student-level variables, however, may be limited (Snijders, 2005). Therefore, in order to include the greatest number of classmates when calculating mean rates of classroom proactive and reactive aggression, we included students with at least 25 min of observed fall behavior ($n = 283$). It would typically require 5 or 6 weeks to collect 25 min of observation. Given the 3-week observer habituation period prior to data collection, 8 weeks of recess participation (two thirds of the total) was probably the minimum time these students were interacting with classmates. All classrooms yielded a minimum of seven observation cases ($M = 8.72$), and fall mean rates were based on a total of 15,077 min of observation time.

Data Analyses

Individuals were nested within classroom, potentially violating the assumption of independence in group variance estimates. Initial unconditional models validated the need for multilevel modeling, and subsequent analyses evaluated competing nested models of the relationships between fall multilevel predictors and spring rates of individual proactive and reactive aggression. The student-level covariates were gender and fall rates of proactive and reactive aggression. Two-grade classrooms accounted for 21.1% of the class sample. Therefore, grade level was treated as a two-level classroom covariate (Grades 3–4 and Grades 5–6). Preliminary analyses found no significant interactions of gender or grade level, confirming their appropriateness as covariates.

Evaluation of three competing models. Predictor variables were grand-mean centered. Models were tested with mixed models (SPSS19) using full-information maximum-likelihood estimation. Models included random effects of the intercept, and competing models were evaluated using Akaike's information criterion (AIC) and likelihood ratio tests.

Results

Descriptive Statistics

Mean total observation times were 49.4 min per student in the fall and 49.5 min in the spring. Correlations between hourly rates of students' proactive aggression, hourly rates of students' reactive aggression, and students' personal beliefs endorsing retaliation for the observed sample are shown in Table 1. Those for the fall are below the diagonal, those for spring above it. Autocorrelations between times are shown on the diagonal. Correlations between students' proactive and reactive aggression were similar to those found in other observational studies (Card & Little, 2006; Polman et al., 2007). Personal beliefs endorsing retaliation showed the greatest stability between fall and spring. Stability (r = .67) and mean endorsement of beliefs (fall M = 0.79, SD = 0.77; spring M = 0.82, SD = 0.81) in the entire sample (N = 536) were virtually identical to those in the observed subsample, and also showed no mean change from fall to spring ($ts < 1.42$, *ns*). Mean rates for students' reactive aggression were significantly higher those for students' proactive aggression in both fall, $t(253) = 6.30$, $p < .01$, $\eta^2 = .22$, and spring, $t(253) = 4.05$, $p < .01$, $\eta^2 = .16$, despite a spring increase in students' proactive, $t(253) = 3.02$, $p < .01$, $\eta^2 = .03$, but not reactive, aggression, $t(253) = 0.08$, *ns*.

Unconditional Models

Unconditional random-effects models were fitted to decompose the variance in observed proactive and reactive aggression into that due to differences across classrooms, and differences across individuals nested within classrooms. In both cases, most of the predictable variance was found at the individual level. Classroom accounted for 17.2% of the variance in students' proactive aggression (intraclass correlation coefficient [ICC] = .172, Wald Z = 2.47, $p < .05$) and 10.1% of the variance in students' reactive aggression (ICC = .101, Wald Z = 1.80, $p < .10$). Since Wald Z is a relatively insensitive test (Hauck & Donner, 1977), the remaining analyses nested subjects in classrooms for both variables.

Table 1
Correlation Matrix and Mean Values for Student Aggression and Personal Beliefs

	Correlations			Spring values	
Measure	PA	RA	Beliefs	*M*	*SD*
PA	**.21**	**.44**	*.17*	1.17[a]	2.10
RA	**.34**	**.32**	.06	1.83[a]	2.72
Beliefs	.07	**.19**	**.72**	.81	.79
Fall values					
M	.75[a]	1.85[a]	.76		
SD	1.29	2.94	.77		

Note. N = 254. Cells below the diagonal provide fall intercorrelations. Cells above the diagonal provide spring intercorrelations. Diagonal cells (underlined) provide fall-spring autocorrelations. Correlations significant at $p < .01$ are in bold. Correlations significant at $p < .05$ are in italics. PA = rate of proactive aggression; RA = rate of reactive aggression; Beliefs = personal beliefs.
[a] rate per hour.

Multilevel Links Between Fall Norms and Students' Aggression

Students' proactive aggression in spring. The first model evaluated the contribution of personal beliefs to hourly rates of students' proactive aggression in the spring after accounting for the contributions of gender, grade level, and fall rates of students' proactive and reactive aggression. Consistent with our hypothesis, personal beliefs predicted high rates of students' proactive aggression (Table 2).

The addition of mean classroom beliefs in Model 2 did not improve model fit. Model 3 added classroom rates of proactive and reactive aggression. The AIC and −2 log likelihood values (difference = 6.40, df = 2, $p < .05$) indicated that Model 3 yielded a better fit. As hypothesized, high rates of classroom reactive aggression in the fall predicted high levels of students' proactive aggression in the spring. In contrast, high classroom rates of proactive aggression in the fall predicted *low* rates of students' proactive aggression in the spring. Even with the addition of classroom aggression, personal beliefs remained significant predictors of students' proactive aggression. Random effects for classroom remained a significant source of variance, Wald Z = 1.97, $p < .05$, after accounting for beliefs and aggression.

Student's reactive aggression in spring. As expected, personal beliefs were not related to students' reactive aggression in the spring after accounting for the covariates (Table 3). As predicted, Model 2 indicated that classroom beliefs predicted later high rates of students' reactive aggression.

The addition of classroom rates of proactive and reactive aggression improved fit over Model 2, as indicated by the AIC and −2 log likelihood values (difference = 8.96, df = 2, $p < .05$). Model 3 indicated that classroom rates of reactive aggression in the fall had a positive relationship to students' reactive aggression in the spring. Classroom rates of proactive aggression, on the contrary, had a negative relationship to students' reactive aggression in the spring. The contribution of classroom beliefs declined to marginal significance with the addition of classroom rates of aggression. The classroom variance remaining after accounting for beliefs and aggression was also not significant, Wald Z = 1.97, *ns*.

Generalizability of results. Analyses indicated that results were not limited by gender or grade. Further, there were no significant interactions between students' aggression rates in the fall and class-level predictor variables.

Discussion

This study adds a cautionary note to investigations of normative influences on aggression, since generalizing across proactive and reactive aggression may prove misleading. Our results demonstrate that normative beliefs contribute to students' proactive and reactive aggression in varied ways that are both theoretically and practically meaningful. Personal beliefs endorsing retaliation pre-

Table 2
Contributions of Fall Aggressive Beliefs and Behavior to Hourly Rates of Students' Proactive Aggression in Spring

Variable	Model 1 Personal beliefs			Model 2 Personal and classroom beliefs			Model 3 Beliefs and classroom behavior		
	Estimate	*SE*	*t* ratio	Estimate	*SE*	*t* ratio	Estimate	SE	*t* ratio
Intercept	1.45	.27	5.31***	1.40	.31	4.47***	1.14	.30	3.75***
Girls	−.42	.24	−1.76†	−.42	.24	−1.76	−.40	.24	−1.68
Grade	.03	.50	.06	.03	.50	.06	.57	.51	1.12
Fall PA rates	.23	.11	2.17*	.23	.11	2.17*	.29	.11	2.63**
Fall RA rates	.17	.04	3.78***	.17	.04	3.78***	.14	.05	3.02**
Personal beliefs	.51	.17	2.91***	.51	.17	2.92***	.53	.17	3.08**
Classroom beliefs				−.91	.79	−.37	−.76	.75	−1.01
Classroom rates of PA							−1.03	.42	−2.43*
Classroom rates of RA							.40	.17	2.34*
Random class effects	.64	.27		.63	.27		.44	.23	
−2 log likelihood			1,026.5			1,026.4			1,020.0
AIC			1,042.5			1,044.4			1,042.0

Note. $N = 254$. PA = rate of proactive aggression; RA = rate of reactive aggression; Belief = personal belief; AIC = Akaike information criterion.
$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

dicted students' proactive aggression, whereas *classroom* beliefs endorsing retaliation predicted students' reactive aggression. Specifically, students whose personal beliefs strongly endorsed retaliation committed one additional act of proactive aggression per hour in the spring, compared to students who did not endorse retaliation. And students in classrooms with strong support for retaliation committed one additional act of reactive aggression every 20 min compared to students in classrooms with less support.

In contrast, the contributions of descriptive classroom norms were similar for proactive and reactive aggression. High classroom rates of proactive aggression in the fall appeared to have an inhibitory influence. Students in classrooms with higher rates of proactive aggression reduced their proactive aggression in the spring by one act per hour, and reactive aggression by one act every 40 min, compared to students in classrooms with lower rates. High classroom rates of reactive aggression made more modest contributions, predicting relative *increases* in students' proactive and reactive aggression by slightly less than one act every two hours.

Functional Relationships Between Beliefs and Aggression

The varied contributions of each type of normative belief are consistent with functional analyses provided by SIP theory and research. While it may seem counterintuitive that personal beliefs endorsing retaliation would predict students' proactive aggression but not reactive aggression, the finding is in line with investigations indicating that preemptive processing during reactive aggression interferes with response evaluation (Step 5 in the SIP model). This diminishes the role of self-regulatory processes (Arsenio & Gold, 2006; Fontaine & Dodge, 2006). Peer actions stemming from normative beliefs (Gasser & Malti, 2012) may have more impact in the heat of the moment because their influence does not rely on response evaluation. Peer encouragement of retaliation

Table 3
Contributions of Fall Aggressive Beliefs and Behavior to Hourly Rates of Students' Reactive Aggression in Spring

Variable	Model 1 Personal beliefs			Model 2 Personal and classroom beliefs			Model 3 Beliefs and classroom behavior		
	Estimate	*SE*	*t* ratio	Estimate	*SE*	*t* ratio	Estimate	*SE*	*t* ratio
Intercept	1.48	.35	4.28***	1.88	.37	5.10***	1.54	.36	4.28***
Girls	.73	.32	2.26*	.77	.32	2.37*	.78	.32	2.42*
Grade	−.04	.45	−.08	−.97	.58	−1.67	−.22	.58	−.38
Fall PA rates	.27	.14	1.87†	.26	.14	1.81	.35	.15	2.37*
Fall RA rates	.23	.06	3.87***	.23	.06	3.92***	.20	.06	3.20**
Personal beliefs	−.11	.30	−.49	−.22	.23	−.92	−.17	.23	−.72
Classroom beliefs				2.13	.92	2.31*	1.60	.87	1.85†
Classroom rates of PA							−1.50	.47	−3.10**
Classroom rates of RA							.44	.20	2.19*
Random class effects	.83	.42		.60	.37		.33	.29	
−2 log likelihood			1,171.8			1,166.9			1,158.0
AIC			1,187.8			1,184.9			1,180.0

Note. $N = 254$. PA = proactive aggression; RA = reactive aggression; AIC = Akaike information criterion.
$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

creates a highly arousing context that impulsive children, those most prone to reactive aggression, may find difficult to resist (Farrell et al., 2010).

The role of personal retaliation beliefs in predicting increased students' proactive aggression is consistent with investigations of self-serving rationales that are employed by young people to justify bullying. Framing actions as a legitimate response to victim behavior (e.g., "He deserved it after what he did to me") is a common element in moral disengagement (Perren et al., 2012), moral justification (de Castro et al., 2012), and selective moral engagement (Hawley & Geldhof, 2012). The convoluted logic that adolescents employ to avoid labeling aggression as bullying by citing *subsequent* retaliatory responses of victims (Boyd, 2014) also supports this possibility. Justifications enable aggressors to bypass moral constraints in themselves and others. Unlike the self-regulatory deficits linked to reactive aggression, such biased justifications (Arsenio & Gold, 2006) do not represent a failure to evaluate aggression (Fontaine & Dodge, 2006). Instead, biases seem attuned to social goals of peer acceptance and social dominance.

Surprisingly, our findings show that students' proactive aggression was only related to concurrent personal beliefs in the spring, while students' reactive aggression was only linked to personal beliefs in the fall. Perhaps justifications for students' proactive aggression become more finely honed in conversations with friends (Caravita, Sijtsema, Rambaran, & Gini, 2014), resulting in greater belief-behavior concordance in the spring. In contrast, individual differences in impulsivity and personal beliefs may contribute less to reactive aggression as the year progresses than the student's status as a victim or nonvictim.

Functional Relationships Between Classroom and Student Rates of Aggression

Students in classrooms with higher rates of proactive aggression in the fall showed spring reductions in both proactive and reactive aggression, relative to those in classrooms with low rates. This might be interpreted in light of presumed function. High classroom rates of proactive aggression may be indicative of a group of students who are competing to establish themselves in dominant positions within a strongly hierarchical class structure. Behavior ecological analyses (Pellegrini, 2008) suggest that when dominance hierarchies stabilize, aggression diminishes. Resignation and fear on the part of subordinate classmates (Camodeca & Goossens, 2005; Craig et al., 2007) may reduce resistance to structural inequities. It is also possible that manipulation and subterfuge on the part of proactive aggressors (Garandeau & Cillessen, 2006; Little et al., 2003; Xie et al., 2002) reduces the likelihood of successful retaliation or competition.

In contrast, students in classrooms with higher rates of reactive aggression in the fall exhibited relatively high rates of proactive and reactive aggression in the spring. This could be indicative of classrooms that were simply more contentious and argumentative than average. In that scenario, reactive aggression would beget more reactive aggression. It might also identify classrooms in which students continued to aggressively resist being subjugated. While moral outrage may result in preemptive processing and ill-considered actions, retaliation against perceived injustice is regarded as a duty in some cultures, even when resistance may be rationally considered to be futile (Frey et al., 2015).

These considerations speak to the role that classroom inequities and injustice can play in provoking reactive aggression. SIP models clearly provide a framework for considering the contributions of moral emotions and beliefs during response evaluation. Arsenio, Adams, and Gold (2009) for example, found that the expectation that unprovoked aggression will lead to positive emotions is uniquely linked to proactive aggression. Research has also examined how beliefs that one has been treated unfairly can be used to legitimize aggression (Perren et al., 2012). There has been surprisingly little attention to how such beliefs might impact in-the-moment emotional reactions that are particularly relevant for reactive aggression. Continued work to integrate work on moral emotions, beliefs, and aggression is needed (Arsenio & Lemerise, 2004; Malti & Latzko, 2010), particularly with respect to processing characteristics identified by SIP models.

Implications for Intervention Efforts and Future Research

The contribution of personal beliefs endorsing retaliation to later students' proactive aggression supports the importance of addressing normative beliefs during intervention efforts. Over time, such efforts may also indirectly reduce reactive aggression by reducing endorsement at the classroom level. One strategy uses attitudinal surveys coupled with personalized feedback aimed at reducing students' tendency to overestimate peer support for aggression. More research is needed to evaluate such approaches (see Henry, 2008 for a discussion of the promise and pitfalls).

Future research is also needed to examine cultural influences regarding retaliation. In "honor cultures" such as those in the southern United States, retaliation is regarded as essential for personal safety and social recognition of manhood. In contrast, "dignity cultures" such as those in New England proscribe personal retaliation as an indicator of poor self-regulation and antisocial behavior. The magnitude of variation suggests that retaliation norms may have considerable real-world significance for the efficacy of intervention practices in different regions or with different populations.

While acknowledging potential cultural differences on this point, it appears that students who are reactively aggressive suffer increasing victimization over time (Salmivalli & Helteenvuori, 2007). Thus, reactive aggression appears inimical to student welfare. Victimization depletes self-regulatory capacity (Baumeister, DeWall, Ciarocco, & Twenge, 2005) in a population that may include a high proportion of impulsive youth. Therefore, intervention programs need to address regulatory skills as well as social norms that support effective nonaggressive responses. Skill remediation may require additional time to achieve reductions in reactive aggression (Frey et al., 2009) or increases in teacher support (Hirschstein, Edstrom, Frey, Snell, & MacKenzie, 2007).

Importantly, interventions need to address issues of justice in order to motivate self-regulation efforts. In line with this thinking, cooperation is higher (Pellegrini, 2008) and aggression lower (Elgar, Craig, Boyce, Morgan, & Vella-Zarb, 2009; Garandeau, Lee, & Salmivalli, 2014) when resources and power are equitably distributed than in a "winner takes all" context. Further, aggression is lower in classrooms whose teachers attempt to mitigate status

differentials between students (Serdiouk, Rodkin, Madill, Logis, & Gest, 2015). Both *degree of stratification* and *hierarchy stability* may influence aggression rates. Ecological analyses indicate that hierarchy stabilization enables social dominants to decrease aggression while still maintaining control (Pellegrini, 2008). The impact of hierarchy stability on victims is unclear. A highly inequitable, yet stable hierarchical structure may provide temporary relief for some, while others remain pariahs (Adler & Adler, 1995). Students may suffer the realization that their personal security is dependent on the whims of others.

A functional analysis integrating information about equality and stability suggests specific hypotheses that are amenable to testing, given sufficient classrooms. Students' proactive aggression is likely to be highest, for example, if destabilization of strongly hierarchical systems creates power struggles. Extreme status inequities raise the stakes associated with being a beneficiary or conversely, someone who bears the cost of injustice. Thus, interventions that appear to be succeeding in strongly hierarchical contexts may unleash reactionary forces aimed at preserving social power (see the meta-analysis of Bettencourt, Charlton, Dorr, & Hume, 2001). A possible result in those classrooms may be a temporary increase or delayed reduction in conflict and aggression. These scenarios highlight the need for intervention studies that measure outcomes more frequently during the course of intervention.

Limitations and Strengths

Several limitations should be noted. We did not measure actual peer influence or perceptions of peers. Although we relied on functional analyses to guide our hypotheses, without knowing individuals' relative influence or the classroom hierarchical structure, our conclusions are necessarily speculative. In addition, high-status classmates are more influential than a random subsample selected for observation (Cohen & Prinstein, 2006). Perceived popularity may be especially important in middle school, when students interact with many students in multiple classrooms. In that case, they often have incomplete information about fellow students, and attending to the behavior exemplified by high-status individuals (Paluck & Shepherd, 2012) may be an efficient way of inferring behavioral norms. Thus, our results may apply primarily to students who remain in one classroom and are quite familiar with the status and aggressive characteristics of their classmates.

An additional caution regarding generalizability is warranted due to the small number and restricted location of schools in this study. We cannot speak to how personal beliefs and actions common at the school-level contribute to later aggression. Nor can we address the possibility that regions that ascribe greater importance to retaliation (Cohen, 2001) might yield different results.

We also do not know how applicable our results are when finer distinctions are made within these specific types of aggression. Bailey and Ostrov (2008) found that associations between young adults' aggressive beliefs and aggressive behavior varied by both aggressive form and function. Proactive relational aggression, for example, might be less effective than proactive physical aggression at inhibiting retaliation. It is also possible that the relationship between aggressive form and function is not orthogonal, such that impulsive, reactive responding is more strongly correlated with overt aggression than with relational aggression (e.g., Frey et al., 2014).

These limitations are juxtaposed against the clarity of precise hourly rates of aggression provided by observers who were blind to hypotheses. Research has been hampered by measurement problems such as the very high correlations between teacher ratings of proactive and reactive aggression common in much past research. This problem has led to the conclusion in two meta-analyses (Card & Little, 2006; Polman et al., 2007) that trained observers are better able to distinguish these two types than are untrained informants. Observer expertise, combined with theoretically coherent links between student- and classroom-level aggressive beliefs and behavior, is a significant strength of this study. Further research into the functional significance of such beliefs and specific types of aggression in school settings will contribute to more effective intervention strategies. Particular benefits may derive from practices that respond to the need for justice that can motivate reactive aggression, and those that provide alternate sources of satisfaction for those motivated by social dominance.

References

Adler, P. A., & Adler, P. (1995). Dynamics of inclusion and exclusion in preadolescent cliques. *Social Psychology Quarterly, 58,* 145–162. http://dx.doi.org/10.2307/2787039

Allen, J. P., Porter, M. R., & McFarland, F. C. (2006). Leaders and followers in adolescent close friendships: Susceptibility to peer influence as a predictor of risky behavior, friendship instability, and depression. *Development and Psychopathology, 18,* 155–172. http://dx.doi.org/10.1017/S0954579406060093

Ang, R. P., Ong, E. Y. L., Lim, J. C. Y., & Lim, E. W. (2010). From narcissistic exploitativeness to bullying behavior: The mediating role of approval-of-aggression beliefs. *Social Development, 19,* 721–735. http://dx.doi.org/10.1111/j.1467-9507.2009.00557.x

Ansary, N. S., Elias, M. J., Greene, M. B., & Green, S. (2015). Guidance for schools selecting antibullying approaches: Translating evidence-based strategies to contemporary implementation realities. *Educational Researcher, 44,* 27–36. http://dx.doi.org/10.3102/0013189X14567534

Arsenio, W. F., Adams, E., & Gold, J. (2009). Social information processing, moral reasoning, and emotion attributions: Relations with adolescents' reactive and proactive aggression. *Child Development, 80,* 1739–1755. http://dx.doi.org/10.1111/j.1467-8624.2009.01365.x

Arsenio, W. F., & Gold, J. (2006). The effects of social injustice and inequality on children's moral judgements and behavior: Towards a theoretical model. *Cognitive Development, 21,* 388–400. http://dx.doi.org/10.1016/j.cogdev.2006.06.005

Arsenio, W. F., & Lemerise, E. A. (2004). Aggression and moral development: Integrating social information processing and moral domain models. *Child Development, 75,* 987–1002. http://dx.doi.org/10.1111/j.1467-8624.2004.00720.x

Bailey, C. A., & Ostrov, J. M. (2008). Differentiating forms and functions of aggression in emerging adults: Associations with hostile attribution biases and normative beliefs. *Journal of Youth and Adolescence, 37,* 713–722. http://dx.doi.org/10.1007/s10964-007-9211-5

Bandura, A., Caprara, G. V., Barbaranelli, C., Pastorelli, C., & Regalia, C. (2001). Sociocognitive self-regulatory mechanisms governing transgressive behavior. *Journal of Personality and Social Psychology, 80,* 125–135.

Baumeister, R. F., DeWall, C. N., Ciarocco, N. J., & Twenge, J. M. (2005). Social exclusion impairs self-regulation. *Journal of Personality and Social Psychology, 88,* 589–604. http://dx.doi.org/10.1037/0022-3514.88.4.589

Bettencourt, B. A., Charlton, K., Dorr, N., & Hume, D. L. (2001). Status differences and in-group bias: A meta-analytic examination of the effects of status stability, status legitimacy, and group permeability. *Psychological Bulletin, 127,* 520–542. http://dx.doi.org/10.1037/0033-2909.127.4.520

Blakemore, S. J., & Mills, K. L. (2014). Is adolescence a sensitive period for sociocultural processing? *Annual Review of Psychology, 65,* 187–207. http://dx.doi.org/10.1146/annurev-psych-010213-115202

Boyd, D. (2014). *It's complicated: The social lives of networked teens.* New Haven, CT: Yale University Press.

Bradshaw, C. P., O'Brennan, L. M., & Sawyer, A. L. (2008). Examining variation in attitudes toward aggressive retaliation and perceptions of safety among bullies, victims, and bully/victims. *Professional School Counseling, 12,* 10–21. http://dx.doi.org/10.5330/PSC.n.2010-12.10

Brendgen, M., Vitaro, F., Barker, E. D., Girard, A., Dionne, G., Tremblay, R. E., & Boivin, M. (2013). Do other people's plights matter? A genetically informed twin study of the role of social context in the link between peer victimization and children's aggression and depression symptoms. *Developmental Psychology, 49,* 327–340. http://dx.doi.org/10.1037/a0025665

Camodeca, M., & Goossens, F. A. (2005). Children's opinions on effective strategies to cope with bullying: The importance of bullying role and perspective. *Educational Research, 47,* 93–105. http://dx.doi.org/10.1080/0013188042000337587

Caravita, S. C. S., Sijtsema, J. J., Rambaran, J. A., & Gini, G. (2014). Peer influences on moral disengagement in late childhood and early adolescence. *Journal of Youth and Adolescence, 43,* 193–207. http://dx.doi.org/10.1007/s10964-013-9953-1

Card, N. A., & Little, T. D. (2006). Proactive and reactive aggression in childhood and adolescence: A meta-analysis of differential relations with psychosocial adjustment. *International Journal of Behavioral Development, 30,* 466–480. http://dx.doi.org/10.1177/0165025406071904

Chang, L. (2004). The role of classroom norms in contextualizing the relations of children's social behaviors to peer acceptance. *Developmental Psychology, 40,* 691–702. http://dx.doi.org/10.1037/0012-1649.40.5.691

Cohen, D. (2001). Cultural variation: Considerations and implications. *Psychological Bulletin, 127,* 451–471. http://dx.doi.org/10.1037/0033-2909.127.4.451

Cohen, G. L., & Prinstein, M. J. (2006). Peer contagion of aggression and health risk behavior among adolescent males: An experimental investigation of effects on public conduct and private attitudes. *Child Development, 77,* 967–983. http://dx.doi.org/10.1111/j.1467-8624.2006.00913.x

Craig, W., Pepler, D., & Blais, J. (2007). Responding to bullying: What works? *School Psychology International, 28,* 465–477. http://dx.doi.org/10.1177/0143034307084136

Crick, N. R., & Dodge, K. A. (1994). A review and reformulation of social information-processing mechanisms in children's social adjustment. *Psychological Bulletin, 115,* 74–101. http://dx.doi.org/10.1037/0033-2909.115.1.74

Crick, N. R., & Grotpeter, J. K. (1995). Relational aggression, gender, and social-psychological adjustment. *Child Development, 66,* 710–722. http://dx.doi.org/10.2307/1131945

de Castro, B. O., Verhulp, E. E., & Runions, K. (2012). Rage and revenge: Highly aggressive boys' explanations for their responses to ambiguous provocation. *European Journal of Developmental Psychology, 9,* 331–350. http://dx.doi.org/10.1080/17405629.2012.680304

Dodge, K. A. (1991). The structure and function of reactive and proactive aggression. In D. J. Pepler & K. H. Rubin (Eds.), *The development and treatment of childhood aggression* (pp. 201–218). Hillsdale, NJ: Erlbaum.

Dodge, K. A., & Coie, J. D. (1987). Social-information-processing factors in reactive and proactive aggression in children's peer groups. *Journal of Personality and Social Psychology, 53,* 1146–1158. http://dx.doi.org/10.1037/0022-3514.53.6.1146

Elgar, F. J., Craig, W., Boyce, W., Morgan, A., & Vella-Zarb, R. (2009). Income inequality and school bullying: Multilevel study of adolescents in 37 countries. *Journal of Adolescent Health, 45,* 351–359. http://dx.doi.org/10.1016/j.jadohealth.2009.04.004

Erdley, C. A., & Asher, S. R. (1996). Children's social goals and self-efficacy perceptions as influences on their responses to ambiguous provocation. *Child Development, 67,* 1329–1344. http://dx.doi.org/10.2307/1131703

Evans, S. C., Fite, P. J., Hendrickson, M. L., Rubens, S. L., & Mages, A. K. (2015). The role of reactive aggression in the link between hyperactive–impulsive behaviors and peer rejection in adolescents. *Child Psychiatry and Human Development, 46,* 903–912. http://dx.doi.org/10.1007/s10578-014-0530-y

Faris, R., & Felmlee, D. (2011). Status struggles network centrality and gender segregation in same-and cross-gender aggression. *American Sociological Review, 76,* 48–73. http://dx.doi.org/10.1177/0003122410396196

Farrell, A. D., Henry, D. B., Schoeny, M. E., Bettencourt, A., & Tolan, P. H. (2010). Normative beliefs and self-efficacy for nonviolence as moderators of peer, school, and parental risk factors for aggression in early adolescence. *Journal of Clinical Child and Adolescent Psychology, 39,* 800–813. http://dx.doi.org/10.1080/15374416.2010.517167

Fontaine, R. G., & Dodge, K. A. (2006). Real-time decision making and aggressive behavior in youth: A heuristic model of response evaluation and decision (RED). *Aggressive Behavior, 32,* 604–624. http://dx.doi.org/10.1002/ab.20150

Fossati, A., Raine, A., Borroni, S., Bizzozero, A., Volpi, E., Santalucia, I., & Maffei, C. (2009). A cross-cultural study of the psychometric properties of the Reactive-Proactive Aggression Questionnaire among Italian nonclinical adolescents. *Psychological Assessment, 21,* 131–135. http://dx.doi.org/10.1037/a0014743

Frey, K. S., Hirschstein, M. K., Edstrom, L. V., & Snell, J. L. (2009). Observed reductions in school bullying, nonbullying aggression, and destructive bystander behavior: A longitudinal evaluation. *Journal of Educational Psychology, 101,* 466–481. http://dx.doi.org/10.1037/a0013839

Frey, K. S., Newman, J. B., & Onyewuenyi, A. C. (2014). Aggressive forms and functions on school playgrounds: Profile variations in interaction styles, bystander actions, and victimization. *The Journal of Early Adolescence, 34,* 285–310. http://dx.doi.org/10.1177/0272431613496638

Frey, K. S., Pearson, C. R., & Cohen, D. (2015). Revenge is seductive, if not sweet: Why friends matter for prevention efforts. *Journal of Applied Developmental Psychology, 37,* 25–35. http://dx.doi.org/10.1016/j.appdev.2014.08.002

Garandeau, C., & Cillessen, A. H. N. (2006). From indirect aggression to invisible aggression: A conceptual view on bullying and peer group manipulation. *Aggression and Violent Behavior, 11,* 612–625. http://dx.doi.org/10.1016/j.avb.2005.08.005

Garandeau, C. F., Lee, I. A., & Salmivalli, C. (2014). Inequality matters: Classroom status hierarchy and adolescents' bullying. *Journal of Youth and Adolescence, 43,* 1123–1133. http://dx.doi.org/10.1007/s10964-013-0040-4

Gasser, L., & Malti, T. (2012). Children's and their friends' moral reasoning: Relations to aggressive behavior. *International Journal of Behavioral Development, 36,* 358–366. http://dx.doi.org/10.1177/0165025412448353

Hauck, W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association, 72,* 851–853.

Hawley, P. H. (2003). Prosocial and coercive configurations of resource control in early adolescence: A case for the well-adapted Machiavellian.

Merrill-Palmer Quarterly, 49, 279–309. http://dx.doi.org/10.1353/mpq.2003.0013

Hawley, P. H., & Geldhof, G. J. (2012). Preschoolers' social dominance, moral cognition, and moral behavior: An evolutionary perspective. *Journal of Experimental Child Psychology, 112,* 18–35. http://dx.doi.org/10.1016/j.jecp.2011.10.004

Hawley, P. H., & Williford, A. (2015). Articulating the theory of bullying intervention programs: Views from social psychology, social work, and organizational science. *Journal of Applied Developmental Psychology, 37,* 3–15. http://dx.doi.org/10.1016/j.appdev.2014.11.006

Henry, D. B. (2008). Changing classroom social settings through attention to norms. In M. Shinn & H. Yoshikawa (Eds.), *Toward positive youth development: Transforming schools and community programs* (pp. 40–57). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195327892.003.0003

Henry, D. B., Dymnicki, A. B., Schoeny, M. E., Meyer, A. L., Martin, N. C., & MVPP. (2013). Middle school students overestimate normative support for aggression and underestimate normative support for nonviolent problem-solving strategies. *Journal of Applied School Psychology, 43,* 433–445.

Henry, D., Guerra, N., Huesmann, R., Tolan, P., VanAcker, R., & Eron, L. (2000). Normative influences on aggression in urban elementary school classrooms. *American Journal of Community Psychology, 28,* 59–81. http://dx.doi.org/10.1023/A:1005142429725

Hirschstein, M. K., Edstrom, L. V., Frey, K. S., Snell, J. L., & MacKenzie, E. P. (2007). Walking the talk in bullying prevention: Teacher implementation variables related to initial impact of the *Steps to Respect* program. *School Psychology Review, 36,* 3–21.

Hubbard, J. A., McAuliffe, M. D., Morrow, M. T., & Romano, L. J. (2010). Reactive and proactive aggression in childhood and adolescence: Precursors, outcomes, processes, experiences, and measurement. *Journal of Personality, 78,* 95–118. http://dx.doi.org/10.1111/j.1467-6494.2009.00610.x

Hubbard, J. A., Smithmyer, C. M., Ramsden, S. R., Parker, E. H., Flanagan, K. D., Dearing, K. F., . . . Simons, R. F. (2002). Observational, physiological, and self-report measures of children's anger: Relations to reactive versus proactive aggression. *Child Development, 73,* 1101–1118. http://dx.doi.org/10.1111/1467-8624.00460

Huesmann, L. R. (1988). An information-processing model for the development of aggression. *Aggressive Behavior, 14,* 13–24. http://dx.doi.org/10.1002/1098-2337(1988)14:1<13::AID-AB2480140104>3.0.CO;2-J

Huesmann, L. R., & Guerra, N. G. (1997). Children's normative beliefs about aggression and aggressive behavior. *Journal of Personality and Social Psychology, 72,* 408–419. http://dx.doi.org/10.1037/0022-3514.72.2.408

Kikas, E., Peets, K., Tropp, K., & Hinn, M. (2009). Associations between verbal reasoning, normative beliefs about aggression, and different forms of aggression. *Journal of Research on Adolescence, 19,* 137–149. http://dx.doi.org/10.1111/j.1532-7795.2009.00586.x

Little, T. D., Henrich, C. C., Jones, S. M., & Hawley, P. H. (2003). Disentangling the 'whys' from the 'whats' of aggressive behavior. *International Journal of Behavioral Development, 27,* 122–133. http://dx.doi.org/10.1080/01650250244000128

Malti, T., & Latzko, B. (2010). Children's moral emotions and moral cognition: Towards an integrative perspective. *New Directions for Child and Adolescent Development, 2010,* 1–10. http://dx.doi.org/10.1002/cd.272

Marsee, M. A., Frick, P. J., Barry, C. T., Kimonis, E. R., Muñoz Centifanti, L. C., & Aucoin, K. J. (2014). Profiles of the forms and functions of self-reported aggression in three adolescent samples. *Development and Psychopathology, 26,* 705–720. http://dx.doi.org/10.1017/S0954579414000339

Mercer, S. H., McMillen, J. S., & DeRosier, M. E. (2009). Predicting change in children's aggression and victimization using classroom-level descriptive norms of aggression and pro-social behavior. *Journal of School Psychology, 47,* 267–289. http://dx.doi.org/10.1016/j.jsp.2009.04.001

Paluck, E. L., & Shepherd, H. (2012). The salience of social referents: A field experiment on collective norms and harassment behavior in a school social network. *Journal of Personality and Social Psychology, 103,* 899–915. http://dx.doi.org/10.1037/a0030015

Pellegrini, A. D. (2008). The roles of aggressive and affiliative behaviors in resource control: A behavioral ecological perspective. *Developmental Review, 28,* 461–487. http://dx.doi.org/10.1016/j.dr.2008.03.001

Pellegrini, A. D., Bartini, M., & Brooks, F. (1999). School bullies, victims, and aggressive victims: Factors relating to group affiliation and victimization in early adolescence. *Journal of Educational Psychology, 91,* 216–224. http://dx.doi.org/10.1037/0022-0663.91.2.216

Pellegrini, A. D., Van Ryzin, M. J., Roseth, C., Bohn-Gettler, C., Dupuis, D., Hickey, M., & Peshkam, A. (2011). Behavioral and social cognitive processes in preschool children's social dominance. *Aggressive Behavior, 37,* 248–257. http://dx.doi.org/10.1002/ab.20385

Perren, S., Gutzwiller-Helfenfinger, E., Malti, T., & Hymel, S. (2012). Moral reasoning and emotion attributions of adolescent bullies, victims, and bully-victims. *British Journal of Developmental Psychology, 30,* 511–530. http://dx.doi.org/10.1111/j.2044-835X.2011.02059.x

Polman, H., de Castro, B. O., Thomaes, S., & van Aken, M. (2009). New directions in measuring reactive and proactive aggression: Validation of a teacher questionnaire. *Journal of Abnormal Child Psychology, 37,* 183–193. http://dx.doi.org/10.1007/s10802-008-9266-0

Polman, H., de Castro, B. O., Koops, W., van Boxtel, H. W., & Merk, W. W. (2007). A meta-analysis of the distinction between reactive and proactive aggression in children and adolescents. *Journal of Abnormal Child Psychology, 35,* 522–535. http://dx.doi.org/10.1007/s10802-007-9109-4

Roseth, C. J., Pellegrini, A. D., Dupuis, D. N., Bohn, C. M., Hickey, M. C., Hilk, C. L., & Peshkam, A. (2011). Preschoolers' bistrategic resource control, reconciliation, and peer regard. *Social Development, 20,* 185–211. http://dx.doi.org/10.1111/j.1467-9507.2010.00579.x

Sackett, G. P. (1977). *Observing behavior: II. Data collection and analysis methods.* Oxford, UK: University Park.

Salmivalli, C., & Helteenvuori, T. (2007). Reactive, but not proactive aggression predicts victimization among boys. *Aggressive Behavior, 33,* 198–206. http://dx.doi.org/10.1002/ab.20210

Sentse, M., Veenstra, R., Kiuru, N., & Salmivalli, C. (2015). A longitudinal multilevel study of individual characteristics and classroom norms in explaining bullying behaviors. *Journal of Abnormal Child Psychology, 43,* 943–955. http://dx.doi.org/10.1007/s10802-014-9949-7

Serdiouk, M., Rodkin, P., Madill, R., Logis, H., & Gest, S. (2015). Rejection and victimization among elementary school children: The buffering role of classroom-level predictors. *Journal of Abnormal Child Psychology, 43,* 5–17. http://dx.doi.org/10.1007/s10802-013-9826-9

Sijtsema, J. J., Veenstra, R., Lindenberg, S., & Salmivalli, C. (2009). Empirical test of bullies' status goals: Assessing direct goals, aggression, and prestige. *Aggressive Behavior, 35,* 57–67. http://dx.doi.org/10.1002/ab.20282

Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1570–1573). Chichester, UK: Wiley. http://dx.doi.org/10.1002/0470013192.bsa492

Stoolmiller, M., Eddy, J. M., & Reid, J. B. (2000). Detecting and describing preventive intervention effects in a universal school-based randomized trial targeting delinquent and violent behavior. *Journal of Consulting and Clinical Psychology, 68,* 296–306.

Thomas, D. E., Bierman, K. L., Powers, C. J., & the Conduct Problems Prevention Research Group. (2011). The influence of classroom aggression and classroom climate on aggressive-disruptive behavior. *Child*

Development, 82, 751–757. http://dx.doi.org/10.1111/j.1467-8624.2011.01586.x

Thomas, K. J., & McGloin, J. M. (2013). A dual-systems approach for understanding differential susceptibility to processes of peer influence. *Criminology: An Interdisciplinary Journal, 51,* 435–474. http://dx.doi.org/10.1111/1745-9125.12010

Ttofi, M. M., & Farrington, D. P. (2011). Effectiveness of school-based programs to reduce bullying: A systematic and meta-analytic review. *Journal of Experimental Criminology, 7,* 27–56. http://dx.doi.org/10.1007/s11292-010-9109-1

Veenstra, R., Lindenberg, S., Munniksma, A., & Dijkstra, J. K. (2010). The complex relation between bullying, victimization, acceptance, and rejection: Giving special attention to status, affection, and sex differences. *Child Development, 81,* 480–486. http://dx.doi.org/10.1111/j.1467-8624.2009.01411.x

Werner, N. E., & Hill, L. G. (2010). Individual and peer group normative beliefs about relational aggression. *Child Development, 81,* 826–836. http://dx.doi.org/10.1111/j.1467-8624.2010.01436.x

Werner, N. E., & Nixon, C. L. (2005). Normative beliefs and relational aggression: An investigation of the cognitive bases of adolescent aggressive behavior. *Journal of Youth and Adolescence, 34,* 229–243. http://dx.doi.org/10.1007/s10964-005-4306-3

Xie, H., Swift, D. J., Cairns, B. D., & Cairns, R. B. (2002). Aggressive behaviors in social interaction and developmental adaptation: A narrative analysis of interpersonal conflicts during early adolescence. *Social Development, 11,* 205–224. http://dx.doi.org/10.1111/1467-9507.00195

Xu, S., & Lorber, M. F. (2014). Interrater agreement statistics with skewed data: Evaluation of alternatives to Cohen's kappa. *Journal of Consulting and Clinical Psychology, 82,* 1219–1227. http://dx.doi.org/10.1037/a0037489

Yeager, D. S., Fong, C. J., Lee, H. Y., & Espelage, D. L. (2015). Declines in efficacy of anti-bullying programs among older adolescents: Theory and a three-level meta-analysis. *Journal of Applied Developmental Psychology, 37,* 36–51. http://dx.doi.org/10.1016/j.appdev.2014.11.005

Received April 28, 2015
Revision received January 28, 2016
Accepted January 29, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 2, 191–207

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000135

What Happens to the Fish's Achievement in a Little Pond? A Simultaneous Analysis of Class-Average Achievement Effects on Achievement and Academic Self-Concept

Franziska Stäbler and Hanna Dumont
German Institute for International Educational Research, Berlin, Germany

Michael Becker
German Institute for International Educational Research, Berlin, Germany, and Leibniz Institute for Science and Mathematics Education, Kiel, Germany

Jürgen Baumert
Max Planck Institute for Human Development, Berlin, Germany

Empirical studies have demonstrated that students who are taught in a group of students with higher average achievement benefit in terms of their achievement. However, there is also evidence showing that being surrounded by high-achieving students has a negative effect on students' academic self-concept, also known as the big-fish–little-pond effect. In view of the reciprocal relationship between achievement and academic self-concept, the present study aims to scrutinize how the average achievement of a class affects students' achievement and academic self-concept, and how that, in turn, affects subsequent achievement and academic self-concept. Using a sample of 6,463 seventh-graders from 285 classes in Germany, multilevel path models showed that the class-average achievement at the beginning of the school year positively affected individual achievement in the middle and at the end of the school year, and negative effects on academic self-concept occurred only at the beginning of Grade 7, but not later in the school year. In addition, mediation analyses revealed that the effects of class-average achievement on students' achievement and academic self-concept at the end of the school year were mediated by midterm achievement, but not by midterm academic self-concept. This pattern was found for mathematics, biology, physics, and English as a foreign language. The results of our study indicate that the consequences for students of belonging to a group of high-achieving students should be analyzed with respect to both academic self-concept and achievement.

Keywords: compositional effects, big-fish–little-pond effect, academic self-concept, reciprocal effects model

Whether and how students are influenced by their class- and schoolmates, is a perpetual topic in education. There is a popular belief that it is better for students to be surrounded by smart and high-achieving peers. And in fact, there are a number of studies showing that the average achievement level of a class or school has a positive effect on subsequent individual achievement over and above students' prior achievement (Burns, & Mason, 2002; Hanushek, Kain, Markman, & Rivkin, 2003; Marks, 2010; Opdenakker, van Damme, de Fraine, van Landeghem, & Onghena, 2002). However, there is also research, in particular from educational psychology, showing that being in a high-achieving class or school can have detrimental effects on other outcomes, such as educational aspirations, academic interests, or academic self-concept (for an overview, see Marsh et al., 2008). Numerous studies on the so-called *big-fish–little-pond effect* (BFLPE; Marsh, 1987) have demonstrated that students with the same achievement level have a lower academic self-concept when they are in a class of high-achieving students than when they are in a class of low-achieving students (for an overview, see Marsh et al., 2008). Being surrounded by high-achieving peers thus has different effects depending on the outcome: It has a positive effect on students' achievement but a negative effect on their academic self-concept. Given that schools should foster both students' achievement as well as their motivational development, it is important to investigate how the average achievement of a group affects students' academic development both in terms of their achievement and their academic self-concept. To our knowledge, a simultaneous analysis of

This article was published Online First June 23, 2016.

Franziska Stäbler and Hanna Dumont, Department of Educational Governance, German Institute for International Educational Research, Berlin, Germany; Michael Becker, Department of Educational Governance, German Institute for International Educational Research, and Department of Educational Research, Leibniz Institute for Science and Mathematics Education, Kiel, Germany; Jürgen Baumert, Center for Educational Research, Max Planck Institute for Human Development, Berlin, Germany.

We are grateful to Benjamin Nagengast and Oliver Lüdtke for valuable comments and feedback on our statistical analyses. Furthermore, we thank Roisin Cronin for editorial assistance.

Correspondence concerning this article should be addressed to Franziska Stäbler, Department of Educational Governance, Deutsches Institut für Internationale Pädagogische Forschung/German Institute for International Educational Research (DIPF), Warschauer Straβe 34-38, D-10243 Berlin, Germany. E-mail: staebler@dipf.de

these two opposite effects of class- or school-average achievement has not been conducted. This is all the more surprising considering that academic self-concept and achievement are reciprocally associated with each other on the individual level. That is, having a high academic self-concept has positive effects on subsequent achievement and vice versa (Marsh & Martin, 2011).

Therefore, the goal of the present study is to scrutinize the interplay between the effect of class-average achievement on achievement and the effect of class-average achievement on academic self-concept. More specifically, using a longitudinal dataset with three measurement points, the study analyzes how the average achievement of a class affects students' achievement and academic self-concept, and how that, in turn, affects subsequent achievement and academic self-concept. Metaphorically speaking, taking into account both achievement development and academic self-concept, is a fish better off swimming in a "big pond" or in a "little pond"?

Before describing our own study in more depth, we describe empirical findings from different research strands that are brought together in our research question: First, we discuss previous research on the positive effects of average achievement on students' individual achievement, which is referred to in the literature using the term compositional effects. Second, we turn to the negative effects of average achievement on students' academic self-concept, known as the BFLPE. Third, we lay out the reciprocal relationship between achievement and academic self-concept on the individual level, which has also been called the reciprocal effects model. Fourth and finally, we present findings from the few empirical studies that have, like ours, looked at both the effects of average achievement on students' individual achievement and academic self-concept.

Positive Effects of Average Achievement on Students' Individual Achievement

Since at least the publication of the Coleman report (Coleman et al., 1966), educational researchers have been investigating whether the composition of a class or school influences the individual development of students. Just as each child is unique, there are a lot of differences between the student bodies of classes or schools. In most cases, students are not randomly assigned to a classroom or a school, which leads to systematic differences between them. On the one hand, the composition of a group of students is determined by the population of the district or neighborhood in which the school is located (i.e., *implicit tracking*; Hallinan, 1994; Trautwein, Köller, Lüdtke, & Baumert, 2005). Moreover, in many educational systems, students are allocated to different school tracks or ability groups according to achievement level (i.e., *explicit tracking or ability grouping*; Ireson & Hallam, 2001; Oakes, 1985; Trautwein et al., 2005). Therefore, schools or classes can be characterized by a certain composition with respect to student characteristics such as achievement, socioeconomic status (SES), or ethnicity (see van Ewijk & Sleegers, 2010). A compositional effect is thus the effect of the school or classroom composition on students' individual achievement over and above the effects of student characteristics at the individual level (Harker & Tymms, 2004). Figure 1 shows the path model of the effect of school/class-average achievement on individual achievement after controlling for prior individual achievement, which is the focus of the present article (for readability, we also refer to this effect as the "compositional effect on achievement").

Indeed, there is empirical evidence that belonging to a group of high achievers has a positive effect on the development of students' individual achievement, whereas being surrounded by low-achieving peers usually has a negative effect (e.g., Hanushek et al., 2003; Marks, 2010). One explanation for such effects concerns the reciprocal influence of students on one another in areas that are associated with achievement, such as motivation or learning effort (Harker & Tymms, 2004). Another mechanism by which compositional effects can be explained is teachers adapting their instruction to the composition of a group of students, for example, by providing cognitively more demanding instruction for high-achieving groups or by covering more subject matter in the same amount of time than they would in a lower-achieving class (Dreeben & Barr, 1988; Harker & Tymms, 2004; Harris, 2010). Additionally, teachers may also have higher expectations for students in high-achieving classes, which can have a positive effect on students' actual performance (Harker & Tymms, 2004; Jussim & Harber, 2005). In general, compositional effects have been found to be larger when looking at the composition of a class rather than that of a school, as the class is the immediate learning environment to which students belong (van Ewijk & Sleegers, 2010).

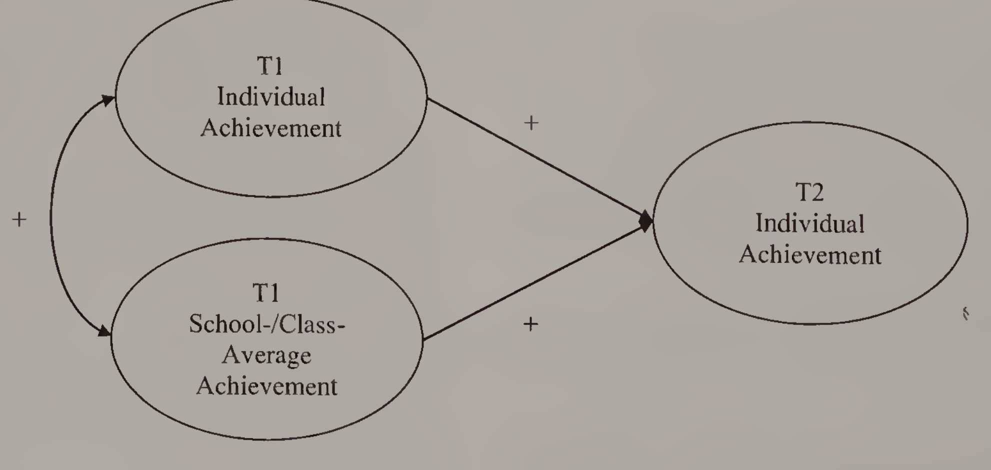


Figure 1. Theoretical path model of the compositional effect of class-average achievement on individual achievement. T1 = Time 1; T2 = Time 2. Plus signs indicate a positive effect; minus signs indicate a negative effect.

Taken together, students in high-achieving classes may benefit from the composition of their classes, whereas students in lower-performing classes are more likely to experience disadvantages with respect to how much they learn. For educational systems with rigid tracking practices, the compositional effect of average achievement may result in a systematic disadvantage for low-achieving students, who are typically allocated to a group with other low-achieving students.

Negative Effects of Average Achievement on Students' Academic Self-Concept

Whereas students in high-achieving classes benefit with respect to their achievement development, there is a large amount of research showing that belonging to a high-achieving group of students has negative consequences for students' academic self-concept (Janssen, Wouters, Huygh, Denies, & Verschueren, 2015; Jonkmann, Becker, Marsh, Lüdtke, & Trautwein, 2012; Liou, 2014; Marsh, 2005; Marsh et al., 2014; Marsh & O'Mara, 2010; Marsh, Trautwein, Lüdtke, Baumert, & Köller, 2007; Nagengast & Marsh, 2011; Roy, Guay, & Valois, 2015; Seaton, Marsh, & Craven, 2010; Wang, 2015; Wouters, de Fraine, Colpin, van Damme, & Verschueren, 2012). This phenomenon is explained by the fact that people compare themselves to others in their reference group (social comparison theory; Festinger, 1954). For students, their classrooms and their schools are their immediate reference groups, which they use for comparisons to form perceptions about their own competencies (Marsh, 1984). Being surrounded by high-achieving students provides more opportunities for "upward" comparisons, which weaken students' academic self-concept. In contrast, belonging to a low-achieving group of students will result in a more positive estimate of a student's competencies due to more "downward" comparisons. This effect is known as the BFLPE, defined as the negative effect of average achievement on students' academic self-concept while controlling for individual achievement (Marsh, 1987). Figure 2 shows the corresponding path model. As the BFLPE describes an effect of average achievement on an individual outcome over and above individual achievement, it can also be considered a compositional effect. However, in line with previous research, we refer to it as BFLPE or describe it as "the effect of class-average achievement on academic self-concept."

A remarkable number of studies have focused on the BFLPE. Just in the past decade, more than 70 articles have been published in leading American Psychological Association journals on the topic. In fact, the BFLPE is probably one of the best researched phenomena in educational psychology. Studies on the BFLPE have, for instance, demonstrated its generalizability across different cultures (Marsh, Kong, & Hau, 2000; Marsh et al., 2012; Nagengast & Marsh, 2011; Seaton, Marsh, & Craven, 2009; Wang, 2015) and its replicability in different grade levels (for a review, see Seaton & Craven, 2011; see also Marsh et al., 2008) and in different subjects (Janssen et al., 2015; Liou, 2014). Moreover, studies have shown that the BFLPE can influence academic self-concept and other academic outcomes, such as achievement, even after graduation from high school (e.g., Marsh & O'Mara, 2010; Marsh et al., 2007).

Typically, the BFLPE is modeled in cross-sectional studies, as Figure 2 shows (for a review, see Marsh et al., 2008). In studies analyzing the BFLPE in a longitudinal framework (e.g., Marsh, Köller, & Baumert, 2001; Marsh et al., 2007), the effect decreases when controlling for students' previous academic self-concept. Whereas in some of the studies, the decreased BFLPE remains relatively high (Köller & Baumert, 2001; Marsh & O'Mara, 2010), in other studies, the remaining effect is small (Marsh et al., 2000; Köller, Trautwein, Lüdtke, & Baumert, 2006), and in some studies, the effect is no longer significant (Marsh et al., 2001; Lüdtke & Köller, 2002).

In sum, we can state that in high-achieving learning environments, students' academic self-concept declines, whereas low-achieving classes or schools can protect students' academic self-concept because they offer fewer opportunities for upward comparisons. This pattern is the reverse of that for the compositional effect on achievement, in which students benefit in the development of their individual achievement in high-achieving classes or schools.

The Reciprocal Relationship Between Achievement and Academic Self-Concept

When investigating the effects of average achievement on students' achievement and their academic self-concept, it is important to bear in mind that achievement and academic self-concept are reciprocally related at the individual level. High achievement leads

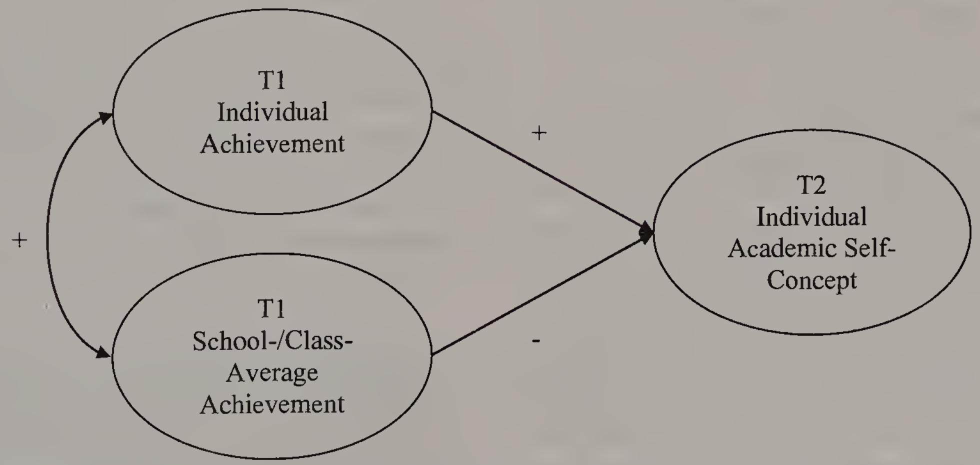


Figure 2. Theoretical path model of the big-fish–little-pond effect. T1 = Time 1; T2 = Time 2. Plus signs indicate a positive effect; minus signs indicate a negative effect.

to a higher academic self-concept (known as the skill-development model; first mentioned by Calsyn & Kenny, 1977; empirically tested by e.g., Skaalvik & Hagtvet, 1990), and a high academic self-concept leads to higher achievement (known as the self-enhancement model; first mentioned by Calsyn & Kenny, 1977; meta-analysis by Valentine, DuBois, & Cooper, 2004). The mutual relationship between achievement and academic self-concept at the individual level, known as the reciprocal effects model (REM; for an early overview, see Byrne, 1984; for a current review, see Marsh & Martin, 2011), is illustrated in Figure 3.

Empirical evidence of the REM has been presented for different grade levels (Helmke & van Aken, 1995; Pinxten, Marsh, de Fraine, Van den Noortgate, & van Damme, 2014; Seaton, Parker, Marsh, Craven, & Yeung, 2014), as well as for different subjects (mathematics, e.g., Pinxten et al., 2014; first language, e.g., Retelsdorf, Köller, & Möller, 2014). Researchers have found the cross-lagged paths of the REM, not only between two measurement points, but also for three and more measurement points (Marsh & O'Mara, 2008; Seaton et al., 2014). With respect to the question of which of the paths is stronger, the findings have been inconsistent. Some studies have found achievement to have stronger effects on academic self-concept (Helmke & van Aken, 1995; van Damme, Opdenakker, de Fraine, & Mertens, 2004), whereas others have shown the opposite (Marsh, Hau, & Kong, 2002; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005). Taken together, when investigating the interplay between the compositional effect on achievement and the BFLPE over time, one must take into account the paths between achievement and academic self-concept at the individual level.

The Relationship Between the Effects of Average Achievement on Achievement and on Academic Self-Concept

Looking at both research on compositional effects on achievement and research on the BFLPE, one can conclude that the composition of a student's class or school in terms of its average achievement is an important factor when understanding a student's academic development. However, depending on the outcome, one may come to opposite conclusions with respect to the question of whether it is beneficial to be in a high-achieving class or school. In order to develop a complete picture of the effects of the average achievement of a class or school on the academic development of students, a simultaneous investigation of both effects over time is needed and, in fact, has been proposed (Chiu, 2012; Dai & Rinn, 2008; Wouters et al., 2012).

Even though a number of studies have modeled average achievement effects on individual achievement and academic self-concept (e.g., Köller & Baumert, 2001[1]; Köller et al., 2006[2]), they have not explicitly addressed the relative importance of the compositional effect on achievement versus the BFLPE. For instance, when they investigated the effects of achievement grouping from Grade 7 to 10, Köller and Baumert (2001) found a significant BFLPE but no compositional effect on achievement when controlling for school track. In another study by Köller et al. (2006) that focused on the interaction between achievement, academic self-concept, and interest between Grades 10 and 12 in the academic track, the authors found significant effects of school-average achievement on individual achievement and on academic self-concept. However, both of these studies investigated the effects of school-average achievement on individual achievement and academic self-concept in separate models. In a study looking at students' achievement and academic self-concept simultaneously, Marsh and O'Mara (2010) found a negative effect of school-average ability on academic self-concept but no longitudinal effects on students' grades or their level of education (which are indicators of students' achievement). Hence, all of these studies have specific limitations with regard to analyzing the relative importance of the composition effect on achievement and the BFLPE. There are further limitations related to their unit of analysis: These studies focus on school composition, but as the class is the most immediate reference group for students (Marsh, Kuyper, Morin, Parker, & Seaton, 2014) this may be the more relevant context to investigate.

There seems to be only one study that addressed both students' individual achievement and their academic self-concept simultaneously when analyzing the effects of average ability at the class-level (Rindermann & Heller, 2005). The authors found that the positive effects on achievement were larger than the negative effects on academic self-concept. However, they only looked at two time points, the sample was not representative as it comprised a small number of classes and focused on special classes for gifted students, and it did not focus on domain-specific achievement or academic self-concept, but on general ability.

The Present Study

To analyze how the class as a learning environment affects students' individual academic development, it is important to investigate the positive effect of class-average achievement on individual achievement, that is, the compositional effect on achievement, and the negative effect of class-average achievement on academic self-concept, that is, the BFLPE, simultaneously. Therefore, the aim of the present study is to investigate the interplay between the compositional effect on achievement and the BFLPE over time using a longitudinal design with three measurement points (Time 1 [T1]: the beginning of Grade 7; Time 2 [T2]: the middle of Grade 7; Time 3 [T3]: the end of Grade 7). More specifically, our research question reads as follows: How does the average achievement of a class at the beginning of Grade 7 affect students' achievement and academic self-concept in the middle of Grade 7, and how does that, in turn, affect achievement and academic self-concept at the end of Grade 7? We approach this research question by first analyzing the compositional effect on achievement, the BFLPE, and the reciprocal relation between individual achievement and academic self-concept separately over the course of the school year. We then bring these different effects together in one model to address our research question (see Figure 4 for the corresponding path model).

Because there is evidence about the compositional effect on achievement, the BFLPE, and the reciprocal effects model in a

[1] Drawing on data from the BIJU study, as we do, Köller and Baumert (2001) analyzed the consequences of attending different schools for students' academic self-concept and achievement in mathematics from Grades 7 to 10.

[2] Drawing on data from the BIJU study, as we do, Köller et al. (2006) analyzed the interplay between individual achievement, academic self-concept, and interest in Grades 10 to 12.

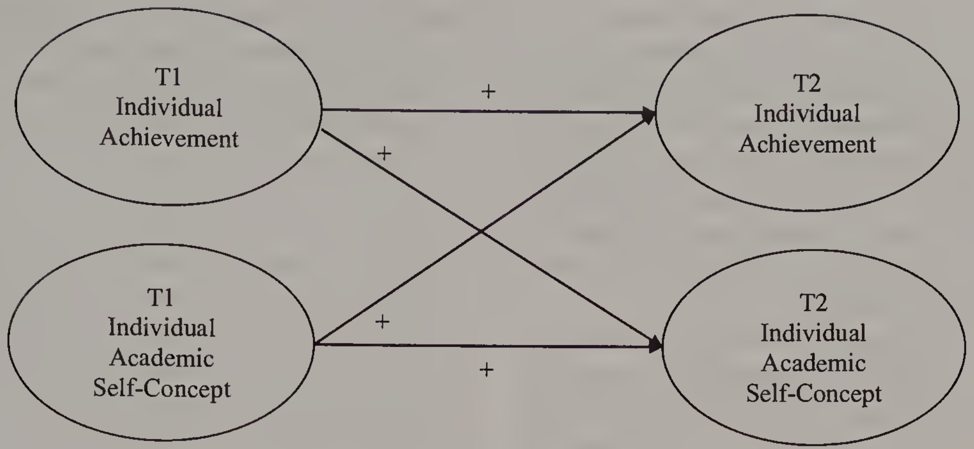


Figure 3. Theoretical path model of the reciprocal effects model. T1 = Time 1; T2 = Time 2. Plus signs indicate a positive effect; minus signs indicate a negative effect.

number of different domains/subjects (for an overview regarding the compositional effect on achievement in different domains, see Dumont et al., 2013; for an overview regarding the BFLPE in different domains, see Marsh et al., 2008; REM in mathematics, e.g., Pinxten et al., 2014; REM in first language, e.g., Retelsdorf et al., 2014), the effects can be assumed to be domain-unspecific effects. In our study, we focus on the academic development of students in the domain of mathematics as an illustrative domain because this area is the focus of a particularly large body of studies on the compositional effect on achievement and the BFLPE (e.g., Harker & Tymms, 2004; Marsh et al., 2007, 2014) and allows us to put our findings into context with previous findings. Mathematics also offers an advantage for analyzing social comparative processes and differences in competencies because, it has been argued, it is especially bound by the curriculum and it is more clearly definable and distinct in terms of its curricular content than other subjects (Gniewosz, 2010). In contrast, reading competencies are relevant in all subjects. To further investigate whether the effects indeed represent domain-unspecific mechanisms and do not emerge solely for the selected domain of mathematics, we replicate the analyses for three other subjects, namely biology, physics, and English as a foreign language.

Method

Sample

The present study draws on data from the longitudinal multicohort study BIJU (Learning Processes, Educational Careers, and Psychosocial Development in Youth and Adolescence), which was conducted by the Max Planck Institute for Human Development in Berlin and begun in the school year 1991/92 with a cohort of seventh-grade students (for more details, see Baumert et al., 1996). Although the study was conducted in four federal states in Germany (Berlin, Saxony–Anhalt [SA], Mecklenburg–Western Pomerania [MWP], North Rhine–Westphalia [NRW]), we only use data from three states, as the survey started at a later measurement point in Berlin. We study data from the first three measurement points of the BIJU study:

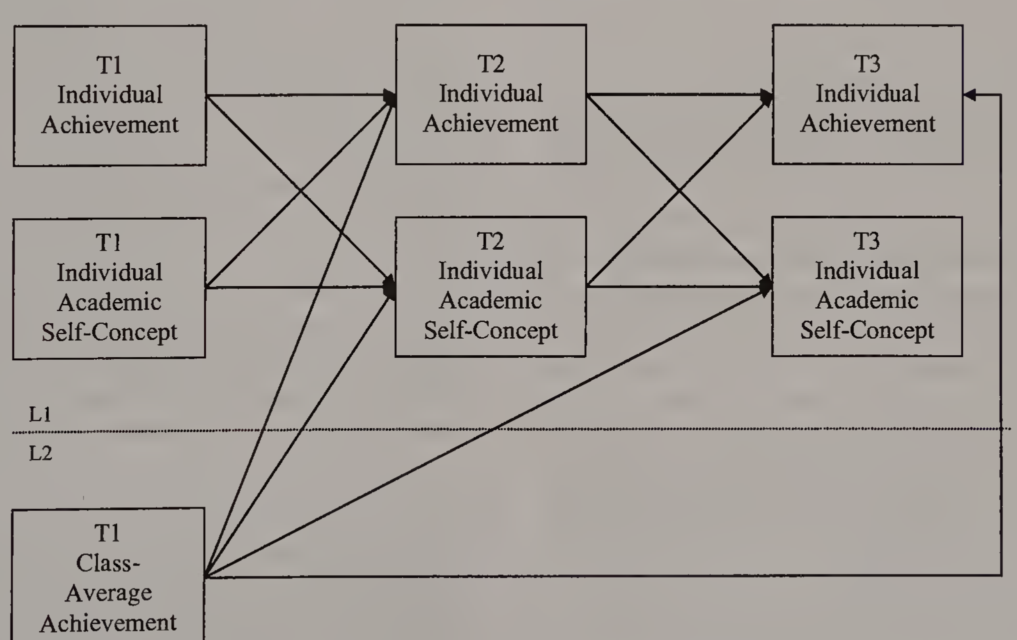


Figure 4. Path model of the present study. For readability, only the paths relevant to our research question are shown here. T1 = Time 1; T2 = Time 2.

the beginning of Grade 7 (T1), halfway through Grade 7 (T2), and the end of Grade 7 (T3). At each measurement point, trained research assistants administered achievement tests on different domain-specific skills and cognitive skills, as well as student questionnaires, on two consecutive school days. The dataset is well suited for our study, as it assesses complete classes and comprises three measurement points for achievement and academic self-concept during a timeframe in which the classroom context does not change.

After removing classes with fewer than 10 students to assure a reliable estimate of class-average achievement, the final sample consisted of 6,463 students from 285 classes at 151 schools (53.2% girls; average age 12.7 years [*SD* = .66]), including 46.8% students from the academic track (59.2% girls; average age 12.6 [*SD* = 0.53]) and 53.2% students from different nonacademic tracks (47.8% girls; average age 12.8 [*SD* = 0.74]). In terms of students' social background, 3.6% of students spoke a language other than German at home, and 47.2% had one or more parents with a college degree. 52.9% of the students attended a school in NRW, 23.3% in MWP, and 23.8% in SA. The three federal states differ in demographic characteristics. The former West-German state of NRW is one of the largest of all German states and had a very high population density in 1991 (514 people per km^2) in contrast to the two former East-German states (MWP: 80 people per km^2, SA: 138 people per km^2; Statistisches Bundesamt, 2010). Moreover, the unemployment rate in the latter two states was higher than in NRW (MWP: 6.6%; SA: 8.3%; NRW: 3.9%; Statistisches Bundesamt, 2010) and the gross domestic product per citizen was lower (US $11,208) than in the average West-German state at the time (US $25,877; Bundesministerium für Wirtschaft und Energie, 2015).

Instruments

Mathematics achievement. Students' mathematical competencies were measured using a curriculum-validated standardized achievement test. Each test comprised approximately 30 items taken from the second international mathematics studies of the International Association for the Evaluation of Educational Achievement (SIMS; Travers & Westbury, 1989), and a study by the Max Planck Institute for Human Development (Baumert, Roeder, Sang, & Schmitz, 1986). The test covered content areas including geometry, algebra, and arithmetic. The anchor test design (Hambleton & Swaminathan, 1989) enabled the estimation of individual achievement test scores on a joint metric for all three assessment rounds via weighted likelihood estimation IRT-modeling (Warm, 1989). More details regarding the content and the scaling of the test can be found in Köller (1998) and Köller, Baumert, and Schnabel (1999). The internal consistency of the test was very high at all three measurement points, that is, Cronbach's alpha was greater than .80 at each point. Intraclass correlation coefficients (ICCs) of .41 (T1), .55 (T2), and .57 (T3) represent a substantial amount of variance in mathematics achievement between classes. This can be seen as an indicator for the grouping of students into different tracks by achievement, a key feature of the German secondary school system. The high ICCs are an important precondition for our study as they imply different learning environments concerning class-average mathematics achievement.

Mathematics self-concept. Students' academic self-concept in mathematics was assessed using five items (see Appendix A) based on Jopt (1978) and Jerusalem (1984). All items required students to use a four-point Likert scale to indicate their agreement (1 = *strongly agree* to 4 = *strongly disagree*). Cronbach's alphas were .83 (T1), .89 (T2), and .90 (T3), showing that the scales were reliable at all three measurement points. This is in line with previous research showing high reliabilities of this scale indicated by Cronbach's alphas larger than .80 (for the mathematics self-concept scales, see also Köller, Daniels, Schnabel, & Baumert, 2000; Möller & Köller, 1995). Moreover, previous research has shown a high construct validity of this scale indicated by high correlations of the mathematics self-concept scales with grades and academic achievement (Baumert, Schnabel, & Lehrke, 1998; Möller & Köller, 1995). The ICCs were .07 (T1), .06 (T2), and .08 (T3). These small sizes of explained variation on the class level are an indicator of the phenomenon that students rate their own competencies in comparison to their class members (see theoretical background).

Control variables. Control variables included gender (with girls as the reference group), parents' socioeconomic status (measured via the highest Treiman Index of the family; the Treiman Index ranges from 0 to 100, with a higher score indicating a higher status), parents' educational background (measured via a dummy variable indicating whether at least one parent had a college degree), language spoken at home (measured via a dummy variable indicating if students did not speak German at home), school track[3] (as a dummy variable indicating whether students were in the academic track), and federal state (as two dummy-coded variables with NRW as the reference category).

The selection of the control variables is based on empirical research on disparities in achievement and academic self-concept. There is evidence for gender differences in both achievement and academic self-concept, depending on the subject. In the case of mathematics, studies revealed higher achievement (Organisation for Economic Co-operation and Development, 2010) and academic self-concept for boys (Marsh & Yeung, 1998; Skaalvik & Skaalvik, 2004). Family background also predicts academic achievement (Organisation for Economic Co-operation and Development, 2010) and academic self-concept (Craven & Marsh, 2005). Furthermore, we control for school track because, in the German secondary school system, class-average achievement is partly confounded by school track. As school tracks differ not only with respect to students' average achievement but also in other aspects such as teacher quality and curriculum (Baumert, Stanat, & Watermann,

[3] The German secondary school system is characterized by an allocation of students into different school tracks after elementary school according to their prior achievement. Even though there are small differences between the federal states concerning type and the number of different school tracks, the main differentiation in all states is between *Gymnasium* as the academic track leading to the *Abitur* (the prerequisite for university entrance) and the remaining non-academic tracks. This is why we created a dummy variable distinguishing between academic and non-academic track.

2006), it is necessary to control for school track in the specific German context. We further control for potential differences between the federal states. Intercorrelations of all variables considered in the analyses are presented in Appendix B.

Statistical Analyses

We analyzed the effect of class-average achievement (CA-ACH) on students' individual achievement (ACH) and academic self-concept (ASC) in mathematics over time using four analytical steps. First, we investigated the compositional effect on achievement. That is, we analyzed whether the average mathematics achievement of classrooms at the beginning of Grade 7 (CA-ACH1) had an effect on students' individual mathematics achievement in the middle of Grade 7 (ACH2) and at the end of Grade 7 (ACH3) after controlling for their mathematics achievement at the beginning (ACH1). In the second step, we analyzed the BFLPE both cross-sectionally (as done in most studies on the BFLPE) for the beginning of Grade 7 and longitudinally in the middle and at the end of Grade 7; that is, the effect of class-average mathematics achievement at the beginning of Grade 7 (CA-ACH1) on mathematics self-concept at three different time points (ASC1, ASC2, ASC3). In the third step, we estimated cross-lagged paths between students' individual mathematics achievement and their mathematics self-concept over the course of all three measurement points. Finally, we brought these different effects together by modeling all paths simultaneously. In this model, we then also estimated mediation effects for class-average mathematics achievement at T1 on individual mathematics achievement at T3, as well as on mathematics self-concept at T3 via mathematics self-concept at T2 and via individual mathematics achievement at T2.

For each step—except the third step, which focused on effects at the individual level only—we specified multilevel models in Mplus 7.11 (Muthén & Muthén, 1998-2010). The mediation analyses were based on the approach suggested by Pituch and Stapleton (2012), which is suited for multilevel mediations in which the effect of a variable at level 2 on a variable at level 1 is mediated by a variable at level 1 (2 – 1 – 1), as is the case in our study.[4] Following this approach, we multiplied the relevant path coefficients using the constraint command in Mplus. We used confidence intervals (CIs) to assess the statistical significance of the indirect effects (CINTERVAL-Output[5]), as they provide a higher test power than common significance parameters (Pituch & Stapleton, 2012).

To account for differences between classes, we grand-mean centered (Enders & Tofighi, 2007) and z-standardized ($M = 0$, $SD = 1$) all continuous variables at the individual level. For ease of interpretation, we also z-standardized class-average mathematics achievement. In all models, we controlled for federal state, school track, gender, parents' socioeconomic status, parents' educational background, and language spoken at home. To address missing data, we used the Full Information Maximum Likelihood procedure, which is implemented in Mplus (FIML; Muthén & Muthén, 1998-2010). This procedure, along with multiple imputation, is considered to be the state-of-the-art method for handling missing data and obtaining unbiased parameter estimates (Graham, 2012). In the present study, between 17.9% and 34.8% of data concerning the mathematics achievement test and the self-report on mathematics self-concept was missing at the three different measurement points. Most data were missing at T3 because some schools could not be resampled. Regarding students who dropped out at T2 or T3, comparative analyses with the remaining students showed no significant differences in their mathematics self-concept at T1. With respect to other variables, there were only small differences in mathematics achievement ($d_{T1T2} = .08$; $d_{T1T3} = .12$), gender ($d_{T1T2} = .02$; $d_{T1T3} = .10$), highest Treiman Index ($d_{T1T2} = .07$; $d_{T1T3} = .20$), language spoken at home ($d_{T1T2} = .10$; $d_{T1T3} = .16$), and school track ($d_{T1T2} = .02$; $d_{T1T3} = .18$).

Results

Compositional Effect on Achievement: The Effect of Class-Average Mathematics Achievement on Students' Individual Mathematics Achievement

First, we modeled the compositional effect on achievement, that is, the influence of class-average math achievement on students' individual math achievement after controlling for previous individual math achievement (see Table 1). Model 1 estimates the compositional effect from T1 to T2, whereas Model 2 estimates the compositional effect from T1 to T3. Model 3 combines Models 1 and 2, that is, Model 3 estimates the effect of class-average math achievement on individual math achievement at T2 and T3 simultaneously while controlling for individual math achievement at both T1 and T2. In line with previous findings, class-average math achievement had a positive effect on students' individual math achievement at both T2 ($\beta = .16$, $p < .001$; Model 1) and T3 ($\beta = .13$, $p < .001$; Model 2). Model 3 reveals that, even when controlling for individual math achievement at T2 and modeling both compositional effects at T2 and T3 simultaneously, both effects were equally high (for T2: $\beta = .15$, $p < .001$; for T3: $\beta = .13$, $p < .01$).

Big-Fish–Little-Pond Effect: The Effect of Class-Average Mathematics Achievement on Students' Mathematics Self-Concept

In the second step, we analyzed the BFLPE, that is the effect of class-average math achievement on students' math self-concept after controlling for students' math achievement (see Table 2). In Model 4, we modeled the BFLPE in a cross-sectional framework by regressing students' math self-concept

[4] The cross-level mediation approach by Pituch and Stapleton (2012) differs from the multilevel structural equation modeling (MSEM) approach suggested by Preacher, Zhang, and Zyphur (2011). Whereas the MSEM approach focuses on mediation at the between level, the approach by Pituch and Stapleton (2012) is suitable for analyzing cross-level effects mediated by a variable at the individual level, which is the case in the present study.

[5] Pituch and Stapleton (2012) recommend the PRODCLIN program for R. The CINTERVAL-Output-command offers an equivalent function for MPlus (source: http://www.statmodel.com/discussion/messages/11/1281.html?1344653748).

Table 1
The Effect of Class-Average Mathematics Achievement on Students' Individual Mathematics Achievement Over the Course of Grade 7

	Model 1	Model 2	Model 3	
	ACH2	ACH3	ACH2	ACH3
Variable	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)
Level 1: Individual predictors				
ACH1	.37*** (.02)	.32*** (.02)	.37*** (.02)	.17*** (.02)
ACH2				.37*** (.02)
Level 2: Class-level predictors				
CA-ACH1	.16*** (.03)	.13** (.04)	.15*** (.03)	.13** (.05)
Residual variance				
Level 1	.35	.38	.35	.33
Level 2	.10	.14	.10	.15

Note. Control variables are gender, parents' socioeconomic status, parents' educational background, and language spoken at home at Level 1, and school track and federal state at Level 2. ACH1 = students' individual mathematics achievement at the beginning of Grade 7; ACH2 = students' individual mathematics achievement in the middle of Grade 7; ACH3 = students' individual mathematics achievement at the end of Grade 7; CA-ACH1 = classroom average mathematics achievement at the beginning of Grade 7.
*$p < .05$. **$p < .01$. ***$p < .001$.

at T1 on class-average math achievement and individual math achievement at T1. We then also investigated the BFLPE in a longitudinal framework. Models 5a and 5b focus on the influence of class-average math achievement at T1 on math self-concept at T2, with Model 5b additionally controlling for math self-concept at T1. Models 6a and 6b specify the effects of class-average math achievement at T1 on math self-concept at T3, with Model 6b additionally controlling for math self-concept at T1.

In line with previous findings, the BFLPE was evident when analyzed in a cross-sectional framework (Model 4). Class-average math achievement had a negative effect of $\beta = -.12$ ($p < .001$) on students' math self-concept after controlling for individual math achievement. This negative effect was also discernible, although somewhat smaller, from T1 to T2 ($\beta = -.08$, $p < .01$; Model 5a), and from T1 to T3 ($\beta = -.10$, $p < .01$; Model 6a). However, when we additionally controlled for previous math self-concept at T1, the effect was no longer statistically significant ($\beta = -.03$, $p = .24$; Model 5b; $\beta = -.06$, $p = .10$; Model 6b).

Reciprocal Effects Model: Cross-Lagged Associations Between Mathematics Achievement and Mathematics Self-Concept

In the third step, we specified the reciprocal relations between students' math achievement and their math self-concept at the individual level. Although Model 7 presents an analysis of the

Table 2
The Effect of Class-Average Mathematics Achievement on Students' Mathematics Self-Concept Over the Course of Grade 7

	Model 4	Model 5		Model 6	
		a	b	a	b
	ASC1	ASC2	ASC2	ASC3	ASC3
Variable	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)
Level 1: Individual predictors					
ASC1			.54*** (.02)		.43*** (.02)
ACH1	.30*** (.02)	.23*** (.02)	.08*** (.02)	.23*** (.02)	.12*** (.02)
Level 2: Class-level predictors					
CA-ACH1	−.12*** (.03)	−.08** (.03)	−.03 (.03)	−.10** (.04)	−.06 (.03)
Residual variance					
Level 1	.83	.87	.63	.87	.72
Level 2	.05	.04	.03	.06	.04

Note. Control variables are gender, parents' socioeconomic status, parents' educational background, and language spoken at home at Level 1, and school track and federal state at Level 2. ASC1 = mathematics self-concept at the beginning of Grade 7; ASC2 = mathematics self-concept in the middle of Grade 7; ASC3 = mathematics self-concept at the end of Grade 7; ACH1 = students' individual mathematics achievement at the beginning of Grade 7; CA-ACH1 = classroom average mathematics achievement at the beginning of Grade 7.
*$p < .05$. **$p < .01$. ***$p < .001$.

cross-lagged paths between students' math achievement and their math self-concept at T1 and T2, Model 8 analyzes the corresponding paths at T1 and T3, as does Model 9 at T2 and T3. Finally, Model 10 analyzes all paths simultaneously.

As Table 3 shows, the four estimated models replicate previous findings showing that students' math achievement and students' math self-concept mutually reinforce each other. The path representing self-enhancement, that is, the path from math self-concept to math achievement, was $\beta = .11$ ($p < .001$) from T1 to T2 (Model 7), $\beta = .09$ ($p < .001$) from T1 to T3 (Model 8), and $\beta = .05$ ($p < .001$) from T2 to T3 (Model 9). When analyzing all paths simultaneously (Model 10), the coefficients dropped in size to $\beta = .03$ ($p < .05$) from T1 to T3 and $\beta = .03$ ($p = .08$) from T2 to T3. The path representing skill-development, that is, the path from math achievement to math self-concept, was $\beta = .08$ ($p < .001$) from T1 to T2 (Model 7), $\beta = .10$ ($p < .001$) from T1 to T3 (Model 8), and $\beta = .11$ ($p < .001$) from T2 to T3 (Model 9). The latter two paths also dropped in size in Model 10 to $\beta = .05$ ($p < .01$) and $\beta = .05$ ($p < .05$). Despite small variations, the self-enhancement and the skill development paths were generally comparable in size, thus providing evidence for the reciprocal effects model.

The Effect of Class-Average Mathematics Achievement on Students' Mathematics Achievement and Mathematics Self-Concept over Time

In the fourth step, we explicitly addressed the interplay between the compositional effect on achievement and the BFLPE over time by estimating the previous models simultaneously over the time period of one school year. In doing so, we modeled the BFLPE in a longitudinal framework. In Model 11, we estimated the effect of class-average math achievement on students' math achievement and math self-concept simultaneously at T2, controlling for their math achievement and math self-concept at T1. In Model 12, we specified the same paths for the time period of T1 to T3. Model 13 includes the paths for all three measurement points.

As Table 4 shows, Models 11 to 13 reveal the same patterns of results as the separate models specified in the previous sections. Controlling for individual math achievement at T1, the positive effect of class-average math achievement at T1 on math achievement was present at T2 ($\beta = .17, p < .001$; Model 11) and at T3 ($\beta = .14$, $p < .01$; Model 12). Even when controlling for individual math achievement at T2 while modeling the compositional effect on achievement at T3 (Model 13), the positive effect of class-average math achievement remained ($\beta = .14$, $p < .01$). Regarding the BFLPE, as in the separate models, there was no statistically significant effect of class-average math achievement at T1 on students' math self-concept at T2 or T3 after controlling for their math self-concept at T1 (coefficients ranging from $\beta = -.06$ to $\beta = -.03$; see Models 11, 12, and 13). The reciprocal relationship between individual math achievement and math self-concept found in the separate models were also found in the full model. That is, the cross-lagged paths were similar in size, and there were only marginal differences between the different time periods. Figure 5 also depicts all theoretically relevant path coefficients from Model 13.

Taken together, the findings show that class-average math achievement has a continuous positive effect on students' individual math achievement throughout Grade 7, but no continuous negative effect on students' math self-concept after taking into account students' math achievement and math self-concept at the beginning of Grade 7. In order to better understand the interplay between the compositional effect on achievement and the BFLPE, we further investigated mediational effects by analyzing whether the effect of class-average math achievement on individual math achievement and math self-concept at the end of Grade 7 was mediated via math achievement and math self-concept in the middle of Grade 7. In line with the nonsignificant BFLPE described above, we found that the effect of class-average math achievement at T1 on math achievement at T3 was not mediated via students' math self-concept at T2 (indirect effect: $\beta = -.00$; $p = .29$; 95% CI [−.00, .00]). The positive effect of class-average math achievement at T1 on

Table 3
Cross-Lagged Associations Between Students' Achievement and Academic Self-Concept Over the Course of Grade 7

	Model 7		Model 8		Model 9		Model 10			
	ASC2	ACH2	ASC3	ACH3	ASC3	ACH3	ASC2	ACH2	ASC3	ACH3
Variable	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)
Level 1: Individual predictors										
ASC1	.54*** (.02)	.11*** (.01)	.45*** (.02)	.09*** (.01)			.54*** (.02)	.11*** (.01)	.22*** (.02)	.03* (.01)
ACH1	.08*** (.02)	.41*** (.02)	.10*** (.02)	.34*** (.02)			.08*** (.02)	.41*** (.02)	.05** (.02)	.16*** (.02)
ASC2					.53*** (.02)	.05*** (.01)			.41*** (.02)	.03 (.02)
ACH2					.11*** (.02)	.53*** (.02)			.05* (.02)	.44*** (.03)
Level 1 residual variance	.65	.43	.76	.51	.69	.45	.65	.43	.66	.43

Note. Control variables are gender, parents' socioeconomic status, parents' educational background, and language spoken at home at Level 1. ASC1 = mathematics self-concept at the beginning of Grade 7; ASC2 = mathematics self-concept in the middle of Grade 7; ASC3 = mathematics self-concept at the end of Grade 7; ACH1 = students' individual mathematics achievement at the beginning of Grade 7; ACH2 = students' individual mathematics achievement in the middle of Grade 7; ACH3 = students' individual mathematics achievement at the end of Grade 7.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 4
The Effect of Class-Average Achievement on Students' Individual Achievement and Academic Self-Concept Over the Course of Grade 7

	Model 11		Model 12		Model 13			
	ASC2	ACH2	ASC3	ACH3	ASC2	ACH2	ASC3	ACH3
Variable	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)
Level 1: Individual predictors								
ASC1	.54*** (.02)	.12*** (.01)	.43*** (.02)	.10*** (.01)	.54*** (.02)	.12*** (.01)	.21*** (.02)	.04** (.01)
ACH1	.08*** (.02)	.34*** (.02)	.11*** (.02)	.29*** (.02)	.08*** (.02)	.34*** (.02)	.06** (.02)	.16*** (.02)
ASC2							.39*** (.02)	.04** (.02)
ACH2							.08** (.03)	.35*** (.02)
Level 2: Class-level predictors								
CA-ACH1	−.03 (.03)	.17*** (.03)	−.06 (.03)	.14** (.04)	−.03 (.02)	.16*** (.03)	−.05 (.03)	.14** (.04)
Residual variance								
Level 1	.63	.34	.72	.37	.63	.34	.62	.33
Level 2	.03	.10	.04	.14	.03	.10	.05	.15

Note. Control variables are gender, parents' socioeconomic status, parents' educational background, and language spoken at home at Level 1, and school track and federal state at Level 2. ASC1 = mathematics self-concept at the beginning of Grade 7; ASC2 = mathematics self-concept in the middle of Grade 7; ASC3 = mathematics self-concept at the end of Grade 7; ACH1 = students' individual mathematics achievement at the beginning of Grade 7; ACH2 = students' individual mathematics achievement in the middle of Grade 7; ACH3 = students' individual mathematics achievement at the end of Grade 7; CA-ACH1 = classroom average mathematics achievement at the beginning of Grade 7.
* $p < .05$. ** $p < .01$. *** $p < .001$.

students' individual math achievement at T3, that is, the compositional effect on achievement, was partially mediated by individual math achievement at T2 (indirect effect: $\beta = .06$, $p < .001$; 95% CI [.03, .08]). Similarly, we found a statistically significant, albeit small, indirect effect of class-average math achievement at T1 on students' math self-concept at T3 via individual math achievement at T2 ($\beta = .01$, $p < .01$; 95% CI [.00, .02]), but not via math self-concept at T2 ($\beta = -.01$, $p = .23$; 95% CI [−.03, .01]).[6] The findings from the mediational analyses thus confirm that the higher the average math achievement of a class at the beginning of Grade 7, the greater the achievement gains of individual students throughout the school year. In contrast, class-average math achievement at T1 did not affect students' math self-concept over the course of Grade 7. In fact, class-average math achievement even had a small positive buffering effect on math self-concept at T3 via students' individual math achievement at T2.

Additional Analyses: Modeling the Cross-Sectional BFLPE

As the previous section described, to look at the effect of class-average math achievement on individual math achievement and math self-concept simultaneously in the full model, we modeled the BFLPE in a longitudinal framework in order to compare both effects. We did not find a statistically significant negative effect of class-average math achievement on students' math self-concept when we controlled for students' achievement throughout Grade 7. However, in the separate models presented above, we did find a statistically significant BFLPE when modeling it cross-sectionally at the beginning of Grade 7, indicating that the negative effect of class-average math achievement on students' math self-concept happened earlier in the school year. Therefore, it is possible that we underestimated the BFLPE when analyzing the compositional effect on achievement and the BFLPE simultaneously in a longitudinal framework. In order to test how the BFLPE observed at the beginning of the school year affected students' academic development throughout Grade 7, we specified two additional mediational models in which we analyzed whether students' math self-concept at T1 mediated the effect of class-average math achievement at T1 on students' math achievement at T3 and their math self-concept at T3.

In fact, we found a small negative indirect effect of $\beta = -.01$ ($p < .01$; 95% CI [−.02, −.01]) from class-average math achievement at T1 on students' math achievement at T3 via their math self-concept at T1. Additionally, we found students' math self-concept at T3 to be negatively affected by class-average math achievement at T1 via their math self-concept at T1 (indirect effect: $\beta = -.01$, $p < .01$; 95% CI [−.02, −.01]). Therefore, we did find the BFLPE at the beginning of Grade 7 to have a negative effect on students' math achievement and their math self-concept at the end of the school year, which we were not able to see in the full model when we modeled the BFLPE longitudinally. However, the indirect effects of class-average math achievement on students' academic development via math self-concept—which were evident even when modeling the BFLPE cross-sectionally—were not as large as the indirect effects via math achievement described in the previous section.

Additional Analyses: Findings for Other Subjects

In order to investigate whether our findings could be replicated for other school subjects, we also analyzed the full model for the subjects of biology, physics, and English as a foreign

[6] The mediational effect of class-average math achievement at T1 on math self-concept at T3 via individual math achievement at T2 may be interpreted as an inconsistent mediation (MacKinnon, 2008) as the direct effect ($\beta = -.05$) was nonsignificant.

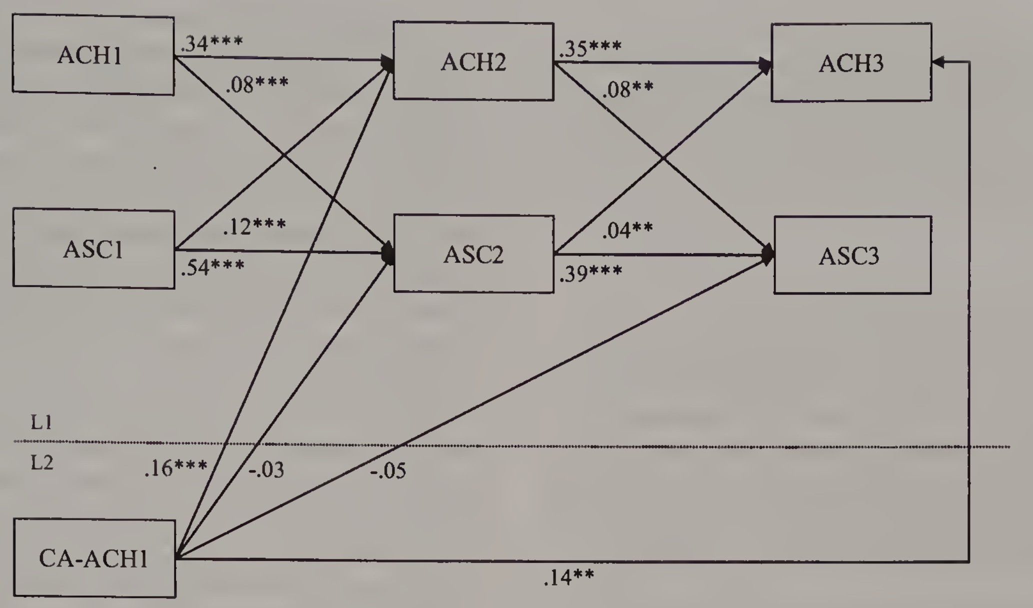


Figure 5. Path model of the relationship between the compositional effect on achievement, the big-fish–little-pond effect, and the reciprocal effects model (Model 13). Only theoretically relevant paths are shown here. Analyses additionally include the estimation of cross-sectional paths between individual achievement and academic self-concept as well as the cross-lagged and auto-correlational paths between Time 1 and Time 3. ASC1 = mathematics self-concept at the beginning of Grade 7; ASC2 = mathematics self-concept in the middle of Grade 7; ASC3 = mathematics self-concept at the end of Grade 7; ACH1 = students' individual mathematics achievement at the beginning of Grade 7; ACH2 = students' individual mathematics achievement in the middle of Grade 7; ACH3 = students' individual mathematics achievement at the end of Grade 7; CA-ACH1 = classroom average mathematics achievement at the beginning of Grade 7; L1 = Level 1 (individual level); L2 = Level 2 (class level). ** $p < .01$. *** $p < .001$.

language, all of which were also assessed in the study.[7] In this section, we summarize the findings from these models with regard to the compositional effect on achievement, the BFLPE, and corresponding mediational effects. In biology, the influence of class-average achievement on individual achievement was $\beta = .17$ ($p < .001$) from T1 to T2, and $\beta = .11$ ($p < .01$) from T1 to T3, and the BFLPE no longer appeared to be significant in the course of the school year (T1 to T2: $\beta = .02$, $p = .54$; T1 to T3: $\beta = .00$, $p = .92$). The positive effect of class-average achievement on individual achievement at T3 was mediated via individual achievement at T2 (indirect effect: $\beta = .07$, $p < .001$; 95% CI [.03, .10]), and the indirect path via academic self-concept at T2 was not significant ($\beta = .00$; $p = .55$; 95% CI [−.00, .00]). Concerning the negative effect of class-average achievement at T1 on academic self-concept at T3, the indirect path coefficients via academic self-concept at T2 ($\beta = .01$, $p = .35$; 95% CI [−.01, .03]) and via individual achievement at T2 ($\beta = .01$, $p = .05$; 95% CI [.00, .02]) were not significant.

In physics, the positive effect of class-average achievement on individual achievement was $\beta = .16$ ($p < .001$) from T1 to T2, and $\beta = .21$ ($p < .001$) from T1 to T3. Class-average achievement had no statistically significant effect on academic self-concept at T2 (T1 to T2: $\beta = .01$, $p = .85$) or at T3 (T1 to T3: $\beta = .01$, $p = .66$). The positive effect on individual achievement at the end of Grade 7 was mediated by individual achievement at T2 (indirect effect: $\beta = .05$, $p < .001$; 95% CI [.03, .07]) but not by academic self-concept at T2 (indirect effect: $\beta = .00$, $p = .84$; 95% CI [−.00, .00]). For the prediction of academic self-concept at T3, there was a small mediation via individual achievement at T2 (indirect effect: $\beta = .01$; $p < .01$; 95% CI [.00, .02]), but not via academic self-concept at T2 (indirect effect: $\beta = .00$; $p = .85$; 95% CI [−.02, .02]).

In English as a foreign language, we found the same pattern. Class-average achievement had a positive effect on individual achievement at T2 ($\beta = .44$, $p < .001$) and at T3 ($\beta = .43$, $p < .001$), but no effect on academic self-concept at T2 ($\beta = -.01$, $p = .75$) or at T3 ($\beta = -.03$, $p = .40$). The positive effect on achievement at T3 was also mediated by individual achievement at T2 (indirect effect: $\beta = .14$, $p < .001$; 95% CI [.11, .18]), whereas the indirect path via academic self-concept at T2 was not significant ($\beta = -.00$, $p = .13$; 95% CI [−.00, .00]). There were no significant indirect paths from class-average achievement at T1 on academic self-concept at T3 via individual achievement at T2 ($\beta =$

[7] The academic achievement in the subjects of biology, physics, and English as a foreign language was measured by standardized tests. Each test in biology was taken from the Second International Science Study IEA (SISS; Rosier & Keeves, 1991), the Lernerfolgstest JT 8 (Zentrum für Schulversuche und Schulentwicklung des Bundesministeriums für Unterricht, Kunst und Sport, 1975), the National Assessment of Educational Progress (NAEP; National Center for Education Statistics, 1989), and the IEA Six-Subject Survey (Walker, 1976). The tests in physics at the three measurement points used items from SISS (Rosier & Keeves, 1991), the Lernerfolgstest JT 8 (Zentrum für Schulversuche und Schulentwicklung des Bundesministeriums für Unterricht, Kunst und Sport, 1975), and from the NAEP (National Center for Education Statistics, 1989). The tests in English as a foreign language at the three measurement points comprised items taken from the IEA Six-Subject Survey (Walker, 1976), the MPI Schulleistungsstudie (Baumert et al., 1986; Edelstein, 1970), and Schrand, Mulch, Portmann, and Stark (1974). The instruments measuring academic self-concept were the same as those used for mathematics; only the terms for the subject were changed.

.01, p = .13; 95% CI [−.01, .04]) or academic self-concept at T2 (β = −.01, p = .13; 95% CI [−.04, .03]). Our findings for the three subjects indicate that, similar to the findings for mathematics, the positive effect of class-average achievement on students' individual achievement was stronger than its negative effect on their academic self-concept.

Discussion

The aim of the present study was to investigate how class-average achievement affects students' individual achievement and their academic self-concept, and how that, in turn, affects subsequent academic self-concept and achievement. We used data collected from students at three measurement points over the course of Grade 7. We focused on mathematics but replicated the findings for other subjects (biology, physics, English as a foreign language) in additional analyses.

The analyses revealed that class-average achievement at the beginning of Grade 7 had a positive effect on students' individual achievement in the middle and at the end of the same school year, controlling for students' achievement at the beginning of Grade 7, thus replicating previous research on the compositional effect on achievement (e.g., Hanushek et al., 2003; Marks, 2010). With respect to the negative effect of class-average achievement on students' academic self-concept after controlling for students' individual achievement—the BFLPE—we only found an effect at the beginning of Grade 7; this effect was not present when we used a longitudinal framework to predict academic self-concept in the middle and at the end of Grade 7 after additionally controlling for academic self-concept at the beginning of Grade 7. This finding is also in line with previous research, which has found strong evidence for the BFLPE in cross-sectional studies but only limited evidence in longitudinal studies (Marsh et al., 2001, 2007). Moreover, we found that students' achievement and academic self-concept were reciprocally related over the course of the school year, which confirms previous findings on the REM (Marsh & O'Mara, 2008; Seaton et al., 2014). When we addressed our research question by analyzing all effects described above simultaneously, we found that the positive effect of class-average achievement on students' achievement was much stronger than the negative effect of class-average achievement on academic self-concept over the course of the school year. In fact, although the compositional effect on achievement was present throughout the entire school year, the BFLPE was only observable at the beginning of the school year. Furthermore, mediation analyses revealed that the effects of class-average achievement on students' achievement and academic self-concept at T3 were mediated by achievement at T2, but not by academic self-concept at T2. That is, the decline in academic self-concept in response to class-average achievement did not result in lower achievement or lower academic self-concept. Taken together, the compositional effect on achievement played a larger role for students' development in our study than did the BFLPE.

Limitations of the Present Study

When interpreting the findings from this study, some limitations must be addressed. First, our study was conducted in the German school system. Even though the German system is particularly well suited for investigating the effects of average achievement because of its tracked system, it would be interesting to investigate whether and how this interplay is apparent in other school systems, particularly those that are less rigidly tracked. In less selective school systems, for example, certain effects may not be as strong.

Second, the present study covered a time period of one school year. This represents only a short period in view of an entire school career, and we cannot draw any conclusions about the interplay between the compositional effect on achievement and the BFLPE over a longer period of time or at a different point in a student's school career.

Third, the data were assessed in 1991/1992. To our knowledge, there was not a more recent dataset that met all the necessary criteria for analyzing our research questions. Even though the compositional effect and the BFLPE are universal context effects that should not have changed over time, a replication of our findings with more recent data would be worthwhile.

Finally, the present study does not make any causal claims as is the case in causal mediation analyses (e.g., Imai, Keele, Tingley, & Yamamoto, 2011; Valeri & VanderWeele, 2013). We do not presume to have identified the mechanisms underlying the observed effects of class-average achievement on individual achievement and academic self-concept, or to have controlled for all potentially confounding factors. Other theoretically plausible variables (e.g., instruction) that were not included in the present study may also be relevant for these relationships. We interpret our results as important descriptive analyses of the relative importance of the compositional effect on achievement and the BFLPE but stress the importance of conducting additional research to provide a more specific causal explanation (for the distinction between description vs. causal explanation, see Foster, 2010).

Theoretical Significance of Our Findings and Implications for Future Research

The marginal role of the BFLPE over the course of a school year might be due to the fact that, after a phase in which students compared themselves with the other students in their class at the beginning of the school year, resulting in changes in their academic self-concept, the relative position of students in their classroom stayed largely the same throughout the year. Hence, we observed no further adjustments in their academic self-concepts. A similar pattern appeared in a study by Wouters et al. (2012), who found that changing from a high-achieving track to a low-achieving track leads to an increase in academic self-concept and a decrease in achievement. Whereas there was a continuous negative effect on academic achievement in the following years for students who changed to a low-achieving track, the development of these students' academic self-concept did not differ from the development of the academic self-concept of students who stayed in the same learning group.

As for the compositional effect on achievement, which was present throughout the whole school year, being surrounded by high-achieving students may have provided a continuously stimulating learning environment for students. More specifically, and taking into account the mechanisms that have been proposed in the literature to explain compositional effects (see review by Dumont et al., 2013; see also van Ewijk & Sleegers, 2010), students may have influenced one another and teachers may have provided

cognitively more demanding instruction for high-achieving groups (e.g., Dreeben & Barr, 1988; Harker & Tymms, 2004; Harris, 2010). Further mechanisms can be assumed in line with social–cognitive theory (Bandura, 1986)—high-achieving students may be role models for their classmates, who imitate the observed learning behavior (i.e., their approaches to solving mathematical problems), or who are influenced implicitly by their motivation to learn (i.e., goal contagion; Aarts, Gollwitzer, & Hassin, 2004).

Based on the cautionary note on causation in the previous section, we can only make assumptions concerning mechanisms underlying the observed effects of class-average achievement on students' individual achievement and their academic self-concept. Therefore, more research is needed in that respect. Understanding the underlying mechanisms is not only theoretically important, but it would also be helpful in order to develop interventions for buffering the negative effects of the BFLPE or supporting students in terms of a more robust and realistic academic self-concept. For instance, recent studies have shown that, for elementary schools, the use of differentiated instruction strategies moderates the BFLPE on academic self-concept (Roy et al., 2015). Similarly, earlier studies showed that when teachers use an individual frame of reference, this has a positive effect on students' academic self-concept (Lüdtke, Köller, Marsh, & Trautwein, 2005) and on their achievement (for an overview, see Mischo & Rheinberg, 1995; Rheinberg & Krug, 1999).

With respect to research on the BFLPE, our study shows that one might come to different conclusions about the importance of this effect for students' academic development when analyzing it in a longitudinal framework. The BFLPE may also be seen in a different light when taking into account the positive effect that class-average achievement has on individual achievement. In terms of the metaphor introduced at the beginning of this article, one may conclude that, despite small losses in academic self-concept, it may be worthwhile for a fish to swim in a "big pond" because of the positive effects on achievement. Therefore, we would like to encourage researchers to not only investigate the negative effects of class-average achievement on students' academic self-concept, but to simultaneously consider other important dimensions of students' academic development, such as individual academic achievement, for which class-average achievement may have positive effects. On a more general note, our study also suggests that research from different strands should be integrated in order to better understand students' development and to arrive at more substantiated findings for practice and policy.

References

Aarts, H., Gollwitzer, P. M., & Hassin, R. R. (2004). Goal contagion: Perceiving is for pursuing. *Journal of Personality and Social Psychology, 87,* 23–37. http://dx.doi.org/10.1037/0022-3514.87.1.23

Bandura, A. (1986). *Social foundations of thought and action: A social–cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Baumert, J., Roeder, P., Gruehn, S., Heyn, S., Köller, O., Rimmele, R., . . . Seipp, B. (1996). Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU) [Educational careers and psychosocial development in adolescence]. In K.-P. Treumann, G. Neubauer, R. Möller, & J. Abel (Eds.), *Methoden und Anwendungen empirischer pädagogischer Forschung* (pp. 170–180). Münster, Germany: Waxmann.

Baumert, J., Roeder, P. M., Sang, F., & Schmitz, B. (1986). Leistungsentwicklung und Ausgleich von Leistungsunterschieden in Gymnasialklassen [Achievement trajectories and reduction of achievement differences]. *Zeitschrift fur Padagogik, 32,* 639–660.

Baumert, J., Schnabel, K.-U., & Lehrke, M. (1998). Learning math in school: Does interest really matter? In L. Hoffmann, A. Krapp, K. A. Renninger, & J. Baumert (Eds.), *Interest and learning: Proceedings of the Seeon Conference on Interest and Gender* (pp. 327–336). Kiel, Germany: Institut für die Pädagogik der Naturwissenschaften an der Universität Kiel.

Baumert, J., Stanat, P., & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus. [School structure and the creation of differential learning and development milieus] In J. Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit/Social inequalities in the educational system: Differential education processes and problems of distributional justice* (pp. 95–188). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften. http://dx.doi.org/10.1007/978-3-531-90082-7_4

Bundesministerium für Wirtschaft und Energie. (Ed.). (2015). *Wirtschaftsdaten Neue Bundesländer* [Economic data new federal states of Germany]. München, Germany: Author.

Burns, R., & Mason, D. A. (2002). Class composition and student achievement in elementary schools. *American Educational Research Journal, 39,* 207–233. http://dx.doi.org/10.3102/00028312039001207

Byrne, B. M. (1984). The general/academic self-concept nomological network: A review of construct validation research. *Review of Educational Research, 54,* 427–456. http://dx.doi.org/10.3102/00346543054003427

Calsyn, R. J., & Kenny, D. A. (1977). Self-concept of ability and perceived evaluation of others: Cause or effect of academic achievement? *Journal of Educational Psychology, 69,* 136–145. http://dx.doi.org/10.1037/0022-0663.69.2.136

Chiu, M.-S. (2012). The internal/external frame of reference model, big-fish–little-pond effect, and combined model for mathematics and science. *Journal of Educational Psychology, 104,* 87–107. http://dx.doi.org/10.1037/a0025734

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, F., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity.* Washington, DC: U. S. Government Printing Office.

Craven, R. G., & Marsh, H. W. (2005). Dreaming futures: An empirical analysis of indigenous Australian students' aspirations, self-concepts, and realities. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *New frontiers for self-research* (pp. 211–232). Greenwich, CT: Information Age.

Dai, D. Y., & Rinn, A. N. (2008). The big-fish–little-pond effect: What do we know and where do we go from here? *Educational Psychology Review, 20,* 283–317. http://dx.doi.org/10.1007/s10648-008-9071-x

Dreeben, R., & Barr, R. (1988). Classroom composition and the design of instruction. *Sociology of Education, 61,* 129–142. http://dx.doi.org/10.2307/2112622

Dumont, H., Neumann, M., Nagy, G., Becker, M., Rose, N., & Trautwein, U. (2013). Class composition effects in non-academic lower secondary school tracks in the state of Baden–Wurttemberg. *Psychologie in Erziehung und Unterricht, 60,* 198–213. http://dx.doi.org/10.2378/peu2013.art16d

Edelstein, W. (1970). Das "Projekt Schulleistung" im Institut für Bildungsforschung in der Max-Planck-Gesellschaft: Berichtet für die Mitarbeiter [The "Projekt Schulleistung" at the Center for Educational Research at Max Planck Institute for Human Development: Report for employees]. *Zeitschrift fur Padagogik, 16,* 517–529.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12,* 121–138. http://dx.doi.org/10.1037/1082-989X.12.2.121

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7,* 117–140. http://dx.doi.org/10.1177/001872675400700202

Foster, E. M. (2010). Causal inference and developmental psychology. *Developmental Psychology, 46,* 1454–1480. http://dx.doi.org/10.1037/a0020204

Gniewosz, B. (2010). Die Konstruktion des akademischen Selbstkonzeptes: Eltern und Zensuren als Informationsquellen [The construction of academic self-concept: Parents and grades as sources of information]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 42,* 133–142. http://dx.doi.org/10.1026/0049-8637/a000014

Graham, J. W. (2012). *Missing data: Analysis and design.* New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-4018-5

Hallinan, M. T. (1994). Tracking: From theory to practice. *Sociology of Education, 67,* 79–91. http://dx.doi.org/10.2307/2112697

Hambleton, R. K., & Swaminathan, H. (1989). *Item response theory: Principles and applications.* Boston, MA: Kluwer-Nijhoff.

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics, 18,* 527–544. http://dx.doi.org/10.1002/jae.741

Harker, R., & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement, 15,* 177–199. http://dx.doi.org/10.1076/sesi.15.2.177.30432

Harris, D. N. (2010). How do school peers influence student educational outcomes? Theory and evidence from economics and other social sciences. *Teachers College Record, 112,* 1163–1197.

Helmke, A., & van Aken, M. A. (1995). The causal ordering of academic achievement and self-concept of ability during elementary school: A longitudinal study. *Journal of Educational Psychology, 87,* 624–637. http://dx.doi.org/10.1037/0022-0663.87.4.624

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *The American Political Science Review, 105,* 765–789. http://dx.doi.org/10.1017/S0003055411000414

Ireson, J., & Hallam, S. (2001). *Ability grouping in education.* London, England: SAGE.

Janssen, R., Wouters, S., Huygh, T., Denies, K., & Verschueren, K. (2015). The effect of peer group performance on the self-concept of reading in a foreign language. *Educational Psychology, 35,* 158–175. http://dx.doi.org/10.1080/01443410.2013.849797

Jerusalem, M. (1984). *Selbstbezogene Kognitionen in schulischen Bezugsgruppen. Eine Längsschnittstudie* [Self-related cognitions in school reference groups: A longitudinal study]. Berlin, Germany: Free University Press.

Jonkmann, K., Becker, M., Marsh, H. W., Lüdtke, O., & Trautwein, U. (2012). Personality traits moderate the big-fish–little-pond effect of academic self-concept. *Learning and Individual Differences, 22,* 736–746. http://dx.doi.org/10.1016/j.lindif.2012.07.020

Jopt, U.-J. (1978). *Selbstkonzept und Ursachenerklärung in der Schule. Zur Attribuierung von Schulleistungen* [Self-concept and causal attributions in schools. On attributions of school achievement]. Bochum, Germany: Kamp.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9,* 131–155. http://dx.doi.org/10.1207/s15327957pspr0902_3

Köller, O. (1998). *Zielorientierungen und schulisches Lernen* [Goal orientations and academic learning]. Münster, Germany: Waxmann.

Köller, O., & Baumert, J. (2001). Leistungsgruppierungen in der Sekundarstufe I. Ihre Konsequenzen für die Mathematikleistung und das mathematische Selbstkonzept der Begabung [Ability grouping at secondary Level 1: Consequences for mathematics achievement and the self-concept of mathematical ability]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 15,* 99–110.

Köller, O., Baumert, J., & Schnabel, K. U. (1999). Wege zur Hochschulreife: Offenheit des Systems und Sicherung vergleichbarer Standards [Ways to higher education entrance qualification: The openness of systems and assurance of comparative standards]. *Zeitschrift für Erziehungswissenschaft, 2,* 385–422.

Köller, O., Daniels, Z., Schnabel, K. U., & Baumert, J. (2000). Kurswahlen von Mädchen und Jungen im Fach Mathematik: Zur Rolle von fachspezifischem Selbstkonzept und Interesse [Course selections of girls and boys in mathematics: The role of academic self-concept and interest]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 14,* 26–37.

Köller, O., Trautwein, U., Lüdtke, O., & Baumert, J. (2006). Zum Zusammenspiel von schulischer Leistung, Selbstkonzept und Interesse in der gymnasialen Oberstufe [On the interplay of academic achievement, self-concept, and interest in upper secondary schools]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 20,* 27–39.

Liou, P.-Y. (2014). Investigation of the big-fish–little-pond effect on students' self-concept of learning mathematics and science in Taiwan: Results from TIMSS 2011. *Asia-Pacific Education Researcher, 23,* 769–778. http://dx.doi.org/10.1007/s40299-013-0152-3

Lüdtke, O., & Köller, O. (2002). Individuelle Bezugsnormorientierung und soziale Vergleiche im Mathematikunterricht: Einfluss unterschiedlicher Referenzrahmen auf das fachspezifische Selbstkonzept der Begabung [Individual reference norm and social comparisons in mathematics classes: The impact of different frames of reference on the domainspecific self-concept of ability]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 34,* 156–166. http://dx.doi.org/10.1026//0049-8637.34.3.156

Lüdtke, O., Köller, O., Marsh, H. W., & Trautwein, U. (2005). Teacher frame of reference and the big-fish–little-pond effect. *Contemporary Educational Psychology, 30,* 263–285. http://dx.doi.org/10.1016/j.cedpsych.2004.10.002

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis.* Mahwah, NJ: Erlbaum.

Marks, G. N. (2010). What aspects of schooling are important? School effects on tertiary entrance performance. *School Effectiveness and School Improvement, 21,* 267–287. http://dx.doi.org/10.1080/09243451003694364

Marsh, H. W. (1984). Self-concept: The application of a frame of reference model to explain paradoxical results. *Australian Journal of Education, 28,* 165–181.

Marsh, H. W. (1987). The big-fish–little-pond effect on academic self-concept. *Journal of Educational Psychology, 79,* 280–295. http://dx.doi.org/10.1037/0022-0663.79.3.280

Marsh, H. W. (2005). Big-fish–little-pond effect on academic self-concept. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 19,* 119–127.

Marsh, H. W., Abduljabbar, A. S., Parker, P. D., Morin, A. J. S., Abdelfattah, F., & Nagengast, B. (2014). The big-fish–little-pond effect in mathematics: A cross-cultural comparison of U.S. and Saudi Arabian TIMSS responses. *Journal of Cross-Cultural Psychology, 45,* 777–804. http://dx.doi.org/10.1177/0022022113519858

Marsh, H. W., Hau, K. T., & Kong, C. K. (2002). Multilevel causal ordering of academic self-concept and achievement: Influence of language of instruction (English compared with Chinese) for Hong Kong students. *American Educational Research Journal, 39,* 727–763. http://dx.doi.org/10.3102/00028312039003727

Marsh, H. W., Köller, O., & Baumert, J. (2001). Reunification of East and West German school systems: Longitudinal multilevel modeling study of the big-fish–little-pond effect on academic self-concept. *American Educational Research Journal, 38,* 321–350. http://dx.doi.org/10.3102/00028312038002321

Marsh, H. W., Kong, C. K., & Hau, K. T. (2000). Longitudinal multilevel models of the big-fish–little-pond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality and Social Psychology, 78,* 337–349. http://dx.doi.org/10.1037/0022-3514.78.2.337

Marsh, H. W., Kuyper, H., Morin, A. J. S., Parker, P. D., & Seaton, M. (2014). Big-fish–little-pond social comparison and local dominance effects: Integrating new statistical models, methodology, design, theory and substantive implications. *Learning and Instruction, 33,* 50–66. http://dx.doi.org/10.1016/j.learninstruc.2014.04.002

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group level effects. *Educational Psychologist, 47,* 106–124. http://dx.doi.org/10.1080/00461520.2012.670488

Marsh, H. W., & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *The British Journal of Educational Psychology, 81,* 59–77. http://dx.doi.org/10.1348/000709910X503501

Marsh, H. W., & O'Mara, A. (2008). Reciprocal effects between academic self-concept, self-esteem, achievement, and attainment over seven adolescent years: Unidimensional and multidimensional perspectives of self-concept. *Personality and Social Psychology Bulletin, 34,* 542–552. http://dx.doi.org/10.1177/0146167207312313

Marsh, H. W., & O'Mara, A. J. (2010). Long-term total negative effects of school-average ability on diverse educational outcomes. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 24,* 51–72.

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K. T., O'Mara, A. J., & Craven, R. G. (2008). The big-fish–little-pond effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review, 20,* 319–350. http://dx.doi.org/10.1007/s10648-008-9075-6

Marsh, H. W., Trautwein, U., Lüdtke, O., Baumert, J., & Köller, O. (2007). The big-fish–little-pond effect: Persistent negative effects of selective high schools on self-concept after graduation. *American Educational Research Journal, 44,* 631–669. http://dx.doi.org/10.3102/0002831207306728

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development, 76,* 397–416. http://dx.doi.org/10.1111/j.1467-8624.2005.00853.x

Marsh, H. W., & Yeung, A. S. (1998). Longitudinal structural equation models of academic self-concept and achievement: Gender differences in the development of math and English constructs. *American Educational Research Journal, 35,* 705–738. http://dx.doi.org/10.3102/00028312035004705

Mischo, C., & Rheinberg, F. (1995). Erziehungsziele von Lehrern und individuelle Bezugsnormen der Leistungsbewertung [Teachers' educational objectives and individualized teacher frame of reference in the context of performance assessments]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 9,* 139–151.

Möller, J., & Köller, O. (1995). Kausalattributionen von Schulleistungen: Reaktive und nicht-reaktive Befragung [Causal attributions of academic achievement: Reactive and non-reactive survey]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 27,* 268–287.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Author.

Nagengast, B., & Marsh, H. W. (2011). The negative effect of school-average ability on science self-concept in the UK, the UK countries, and the world: The big-fish–little-pond effect for PISA 2006. *Educational Psychology, 31,* 629–656. http://dx.doi.org/10.1080/01443410.2011.586416

National Center for Education Statistics (Ed.), (1989). *Report of the NAEP, Technical Review Panel on the 1986 Reading Anomaly, the Accuracy of NAEP trends, and Iissues raised by state-level NAEP comparisons* (Tech. Rep. No. CS 89-499). Washington, DC: US Department of Education, Office of Educational Research and Improvement.

Oakes, J. (1985). *Keeping track: How schools structure inequality.* Binghamton, NY: Vail-Ballou Press.

Organisation for Economic Co-operation and Development. (Ed.). (2010). *PISA 2009. What students know and can do. Student performance in reading, mathematics and science.* Paris, France: Author.

Opdenakker, M.-C., van Damme, J., de Fraine, B., van Landeghem, G., & Onghena, P. (2002). The effect of schools and dasses on mathematics achievement. *School Effectiveness and School Improvement, 13,* 399–427. http://dx.doi.org/10.1076/sesi.13.4.399.10283

Pinxten, M., Marsh, H. W., De Fraine, B., Van Den Noortgate, W., & Van Damme, J. (2014). Enjoying mathematics or feeling competent in mathematics? Reciprocal effects on mathematics achievement and perceived math effort expenditure. *The British Journal of Educational Psychology, 84,* 152–174. http://dx.doi.org/10.1111/bjep.12028

Pituch, K. A., & Stapleton, L. M. (2012). Distinguishing between cross- and cluster-level mediation processes in the cluster randomized trial. *Sociological Methods & Research, 41,* 630–670. http://dx.doi.org/10.1177/0049124112460380

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling, 18,* 161–182. http://dx.doi.org/10.1080/10705511.2011.557329

Retelsdorf, J., Köller, O., & Möller, J. (2014). Reading achievement and reading self-concept: Testing the reciprocal effects model. *Learning and Instruction, 29,* 21–30. http://dx.doi.org/10.1016/j.learninstruc.2013.07.004

Rheinberg, F., & Krug, S. (Eds.). (1999). *Motivationsförderung im Schulalltag* [Strengthening motivation in everyday school life]. Göttingen, Germany: Hogrefe.

Rindermann, H., & Heller, K. A. (2005). The benefit of gifted classes and talent schools for developing students' competences and enhancing academic self-concept. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology, 19,* 133–136.

Rosier, M., & Keeves, J. (Eds.). (1991). *The IEA study of science* (Vol. 1–3). Oxford, England: Pergamon Press.

Roy, A., Guay, F., & Valois, P. (2015). The big-fish–little-pond effect on academic self-concept: The moderating role of differentiated instruction and individual achievement. *Learning and Individual Differences, 42,* 110–116. http://dx.doi.org/10.1016/j.lindif.2015.07.009

Schrand, H., Mulch, D., Portmann, R., & Stark, G. (Eds.). (1974). *Englisch-Einstufungstest 7.* Weinheim, Germany: Beltz.

Seaton, M., & Craven, R. G. (2011). From frogs to fish: The 'big-fish–little-pond' effect then and now. *Psiholoska Obzorja/Horizons of Psychology, 20,* 107–120.

Seaton, M., Marsh, H. W., & Craven, R. G. (2009). Earning its place as a pan-human theory: Universality of the big-fish–little-pond effect across 41 culturally and economically diverse countries. *Journal of Educational Psychology, 101,* 403–419. http://dx.doi.org/10.1037/a0013838

Seaton, M., Marsh, H. W., & Craven, R. G. (2010). Big-fish–little-pond effect: Generalizability and moderation—Two sides of the same coin. *American Educational Research Journal, 47,* 390–433. http://dx.doi.org/10.3102/0002831209350493

Seaton, M., Parker, P., Marsh, H. W., Craven, R. G., & Yeung, A. S. (2014). The reciprocal relations between self-concept, motivation and achievement: Juxtaposing academic self-concept and achievement goal orientations for mathematics success. *Educational Psychology, 34,* 49–72. http://dx.doi.org/10.1080/01443410.2013.825232

Skaalvik, E. M., & Hagtvet, K. A. (1990). Academic achievement and self-concept: An analysis of causal predominance in a developmental

perspective. *Journal of Personality and Social Psychology, 58,* 292–307. http://dx.doi.org/10.1037/0022-3514.58.2.292

Skaalvik, S., & Skaalvik, E. M. (2004). Gender differences in math and verbal self-concept, performance expectations, and motivation. *Sex Roles, 50,* 241–252. http://dx.doi.org/10.1023/B:SERS.0000015555.40976.e6

Statistisches Bundesamt (Ed.), (2010). *20 Jahre Deutsche Einheit: Wunsch oder Wirklichkeit* [20 years of German unity: Desire or reality]. Berlin, Germany: Statistisches Bundesamt.

Trautwein, U., Köller, O., Lüdtke, O., & Baumert, J. (2005). Student tracking and the powerful effects of opt-in courses on self-concept: Reflected-glory effects do exist after all. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *New frontiers for self-research* (pp. 307–327). Greenwich, CT: Information Age.

Travers, K., & Westbury, I. (Eds.). (1989). *The IEA study of mathematics* (Vol. 1–3). Oxford, England: Pergamon Press.

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist, 39,* 111–133. http://dx.doi.org/10.1207/s15326985ep3902_3

Valeri, L., & Vanderweele, T. J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods, 18,* 137–150. http://dx.doi.org/10.1037/a0031034

van Damme, J., Opdenakker, M.-C., de Fraine, B., & Mertens, W. (2004). Academic self-concept and academic achievement: Cause and effect. In H. W. Marsh, J. Baumert, G. E. Richards, & U. Trautwein (]Eds.), *Third International Biennial SELF Research Conference in Berlin* [CD-ROM]. Berlin, Germany: SELF Research Conference.

van Ewijk, R., & Sleegers, P. (2010). Peer ethnicity and achievement: A meta-analysis into the compositional effect. *School Effectiveness and School Improvement, 21,* 237–265. http://dx.doi.org/10.1080/09243451003612671

Walker, D. (Ed.), (1976). *IEA Six-Subject Survey: An empirical study of education in twenty-one countries.* New York, NY: Wiley.

Wang, Z. (2015). Examining big-fish–little-pond-effects across 49 countries: A multilevel latent variable modelling approach. *Educational Psychology, 35,* 228–251. http://dx.doi.org/10.1080/01443410.2013.827155

Warm, T. A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika, 54,* 427–450. http://dx.doi.org/10.1007/BF02294627

Wouters, S., de Fraine, B., Colpin, H., van Damme, J., & Verschueren, K. (2012). The effect of track changes on the development of academic self-concept in high school: A dynamic test of the big-fish–little-pond effect. *Journal of Educational Psychology, 104,* 793–805. http://dx.doi.org/10.1037/a0027732

Zentrum für Schulversuche und Schulentwicklung des Bundesministeriums für Unterricht. Kunst und Sport (Ed.). (1975). *Lernerfolgstest JT 8* [Test for Learning Success JT 8]. Graz, Austria: Bundesministeriums für Unterricht, Kunst und Sport.

Appendix A

Mathematics Self-Concept Items

- I would much prefer math if it weren't so hard. (1 = *strongly agree* to 4 = *strongly disagree*)
- Although I make a real effort, math seems to be harder for me than for my fellow students. (1 = *strongly agree* to 4 = *strongly disagree*)
- Nobody's perfect, but I'm just not good at math. (1 = *strongly agree* to 4 = *strongly disagree*)
- Some topics in math are just so hard that I know from the start I'll never understand them. (1 = *strongly agree* to 4 = *strongly disagree*)
- Math just isn't my Thing. (1 = *strongly Agree* to 4 = *strongly Disagree*)

(*Appendices continue*)

Appendix B

Intercorrelations of All Variables Considered in the Analyses

	ASC			ACH			CA−ACH	Control variables							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. ASC1	—														
2. ASC2	.58***	—													
3. ASC3	.47***	.54***	—												
4. ACH1	.25***	.20***	.21***	—											
5. ACH2	.22***	.20***	.20***	.65***	—										
6. ACH3	.18***	.15***	.22***	.56***	.71***	—									
7. CA-ACH1	.10***	.06***	.09***	.66***	.62***	.58***	—								
8. Sex	.21***	.20***	.11****	.06***	.00	−.07***	−.15*	—							
9. Age	−.03*	−.03*	−.04**	−.15***	−.17***	−.14***	−.47***	.08***	—						
10. Parents' highest Treiman Index	.03*	.03*	.02	.03	−.01	.02	.10	.06***	−.01	—					
11. Parents' educational background	.04*	.03	.02	.17***	.21***	.22***	.62***	.05***	−.07***	.03	—				
12. Language spoken at home	.00	.00	−.04*	−.10***	−.11***	−.11***	−.33***	.04*	.12***	.03	−.06***	—			
13. School track	.08***	.05***	.10***	.51***	.61***	.60***	.77***	−.11***	−.19***	−.10***	.32***	−.10***	—		
14. Federal state Dummy MWP	−.06***	−.08***	−.04*	.08***	.19***	.14***	.09	−.02	−.07***	−.04***	.10***	−.06***	.19***	—	
15. Federal state Dummy SA	−.03*	.00	−.00	.15***	.29***	.26***	.24***	−.02	−.07***	.03	.15***	−.07***	.19***	−.31***	—

Note. Intercorrelations of class-average mathematics achievement at T1 and control variables are estimated at Level 2. Federal state Dummy: MWP = Mecklenburg–Western Pomerania, SA = Saxony–Anhalt; ASC1 = mathematics self-concept at the beginning of Grade 7; ASC2 = mathematics self-concept in the middle of Grade 7; ASC3 = mathematics self-concept at the end of Grade 7; ACH1 = students' individual mathematics achievement at the beginning of Grade 7; ACH2 = students' individual mathematics achievement in the middle of Grade 7; ACH3 = students' individual mathematics achievement at the end of Grade 7; CA-ACH1 = classroom average mathematics achievement at the beginning of Grade 7.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Received June 17, 2015
Revision received April 12, 2016
Accepted April 15, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 2, 208–219

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000130

Homework and Achievement: Using Smartpen Technology to Find the Connection

Kevin Rawson and Thomas F. Stahovich
University of California, Riverside

Richard E. Mayer
University of California, Santa Barbara

There is a long history of research efforts aimed at understanding the relationship between homework activity and academic achievement. While some self-report inventories involving homework activity have been useful for predicting academic performance, self-reported measures may be limited or even problematic. Here, we employ a novel method for accurately measuring students' homework activity using smartpen technology. Three cohorts of engineering students in an undergraduate statics course used smartpens to complete their homework problems, thus producing records of their work in the form of timestamped digitized pen strokes. Consistent with the time-on-task hypothesis, there was a strong and consistent positive correlation between course grade and time doing homework as measured by smartpen technology ($r = .44$), but not between course grade and self-reported time doing homework ($r = -.16$). Consistent with an updated version of the time-on-task hypothesis, there was a strong correlation between measures of the quality of time spent on homework problems (such as the proportion of ink produced for homework within 24 hr of the deadline) and course grade ($r = -.32$), and between writing activity (such as the total number of pen strokes on homework) and course grade ($r = .49$). Overall, smartpen technology allowed a fine-grained test of the idea that productive use of homework time is related to course grade.

Keywords: homework, time-on-task, educational technology, data mining

Homework is defined as "tasks assigned to students by school teachers that are meant to be carried out during non-school hours" (Cooper, 1989, p. 7). Homework has the potential to improve academic learning, perhaps by extending time to learn beyond the classroom and priming active cognitive processing for learning (Cooper, 1989, 2001; Mayer, 2011). Assigning homework problems to be solved by students outside of class time is a common practice in college courses in engineering, mathematics, and science. The goal of the present study is to determine how students' problem-solving activity on homework is related to their course grade in introductory-level engineering courses.

Smartpen Technology

Suppose a teacher assigns homework problems for students to work on each week. How can we know the degree to which students engage with the homework assignment? We could ask them to report how much time (or effort) they put into the homework assignments, but self-reported measures can be problematic. Instead, imagine that a teacher could assign homework problems to students and be able to monitor the student's homework activity at any time and any place, even outside of class. In short, suppose we had a way to know when a student was working on a homework assignment and we were able to record every pen stroke a student made while working on a handwritten assignment. This level of rich data mining of student handwritten homework activity employed in the current study is enabled by the use of newly developed smartpen technology that accomplishes this goal (Herold, Stahovich, Lin, & Calfee, 2011).

Rationale

Researchers have long sought to understand the role of study activities (including homework activities) in academic achievement. For example, Jones and Ruch (1928) examined the relationship between the amount of time spent studying and first semester grade point average. More recently, Credé and Kuncel (2008) conducted a meta-analysis of 10 study habit, skill, and attitude inventories and found that they had incremental validity in predicting academic performance.

Much of this work relies on surveys and students' self-reports of study habits, which may limit the reliability. For example, Schuman, Walsh, Olson, and Etheridge (1985) found little relation between study time and grades, and attributed this to "the possible invalidity of student reports of their own studying" (p. 961). Blumner and Richards (1997) found that a study habit inventory

This article was published Online First June 2, 2016.

Kevin Rawson and Thomas F. Stahovich, Department of Mechanical Engineering, Bourns College of Engineering, University of California, Riverside; Richard E. Mayer, Department of Psychological and Brain Sciences, University of California, Santa Barbara.

This material is based upon work supported by the National Science Foundation under Award Numbers 0935239 and 1432820. Livescribe, Inc. provided some of the materials used in the project and Microsoft provided access to computing resources used for data mining.

Correspondence concerning this article should be addressed to Thomas F. Stahovich, Department of Mechanical Engineering, Bourns College of Engineering, University of California, A349 Bourns Hall, Riverside, CA 92521. E-mail: stahov@engr.ucr.edu

was useful for differentiating between high- and low-performing students. However, the authors concluded that: "It will be necessary to directly observe students in the act of studying. Only in this manner can it be determined that students actually do what they say in response to such an inventory" (p. 132).

In our present work, we take up this challenge, and use Livescribe Smartpens™ to measure students' homework activity. These devices have an integrated camera and are used with dot-patterned paper. They serve the same function as a traditional ink pen and also record the work as timestamped pen strokes. We conducted studies in three offerings of a sophomore-level undergraduate engineering course in statics. Students in these courses completed their homework assignments using the smartpens, thus producing records of the work in the form of timestamped digitized pen strokes.

Homework

There is encouraging evidence—much dating from the 1980s (Keith, 1982)—for the educational value of homework (Cooper, Robinson, & Patall, 2006; Hattie, 2009; Xu, 2013). At the grossest level, Hattie (2009) reported an average effect size of d = .29 favoring homework, based on five meta-analyses involving 295 experimental tests and over 100,000 students. In another review of research on the relation between homework and achievement, Cooper, Robinson, and Patall (2006) found a weighted average correlation of r = .24 based on 69 separate correlations. Importantly, the research team found the positive correlation between homework and achievement was greater for older students (e.g., high school students) than for younger students (e.g., elementary school students).

Although early research focused on the quantity of homework activity (such as the reported time spent on homework), Xu (2013) has proposed that the next step in research on homework is to more carefully examine the quality of homework activity—including the learner's effort and activity. A methodological obstacle to determining the relation between homework and achievement is that much of the existing research is based on students' self-reported time (or effort) on homework rather than on their actual activity. A related methodological obstacle is that the focus is on what homework is assigned by teachers rather than on what is done by students as they work on their homework.

The present study overcomes these challenges by employing a computer-based technology for tracking the details of students' homework activity in real time using smartpens. This technology provides a level of detail about what students are doing and when they are doing it that is not possible in classic research on homework. Thus, this technology-enhanced system provides data for an updated examination of the connection between homework and achievement.

Theory and Predictions

The amount of time that students choose to give to a task can be considered a measure of student engagement (Hattie, 2009; van Gog, 2013). Student engagement during learning is at the heart of theories of meaningful learning such as cognitive load theory (Sweller, Ayres, & Kalyuga, 2011) and the cognitive theory of multimedia learning (Mayer, 2009, 2014), and theories of academic motivation such as self-efficacy theory (Schunk & Pajares, 2009) and attribution theory (Graham & Williams, 2009). Figure 1 shows the proposed causes and consequences of student engagement during learning. In terms of what causes students to exert effort, the left side of Figure 1 proposes that instructional features (such as interactivity and personalization) and student characteristics (such as self-efficacy and interest) can prime the level of student effort during learning. A major task of research on instructional design is to identify instructional features that cause the learner to exert effort to learn, and a major task of research on academic motivation is to identify motivational beliefs that cause the learner to exert effort to learn. In terms of what are the consequences of students engagement, the right side of Figure 1 shows that effort to learn can lead to better learning outcomes as indicated by measures of achievement.

According to this basic model of academic learning, engagement (as indicated by the amount of time that students allocate to a task) is a mechanism affecting learning outcomes (as indicated by achievement). Our focus in the current study is on the relation between time on a study task (i.e., doing homework assignments) and grades in a college course. Thus, our focus is on a crucial link in a model of academic learning. A major new contribution is a more detailed measurement of student engagement on a study activity (i.e., doing handwritten homework assignments) than has been previously available.

Our predictions are based on the time-on-task hypothesis (Hattie, 2009; van Gog, 2013), which holds that learning new material is related to the amount of time a student is effortfully engaged in a productive learning activity. Productive learning activities are those that cause the student to attend to relevant material, mentally organize it, and relate it with relevant prior knowledge (Mayer, 2009, 2014). Spending time on homework is one way to increase productive learning time beyond the school day.

Time-on-task—defined as the amount of time a student spends engaged in an academic task—can be "counted among the most important factors affecting student learning and achievement" (van Gog, 2013, p. 432). Rooted in Ebbinghaus' (1885/1964) classic studies on verbal learning which showed that time spent studying a word list is related to the amount learned, time-on-task has been recognized as a potentially important variable in academic learning since the 1960s (Berliner, 1991; Carroll, 1963; van Gog, 2013). In a review of meta-analyses, Hattie (2009) found an average effect size of d = .38 for time-on-task based on four meta-analyses examining 136 experimental comparisons.

Over the years, the concept of time-on-task has evolved to reflect a focus on *engaged learning time*—time in which the learner is actively exerting effort on a task—rather than *allocated learning time*—time in which the instructor provides opportunities

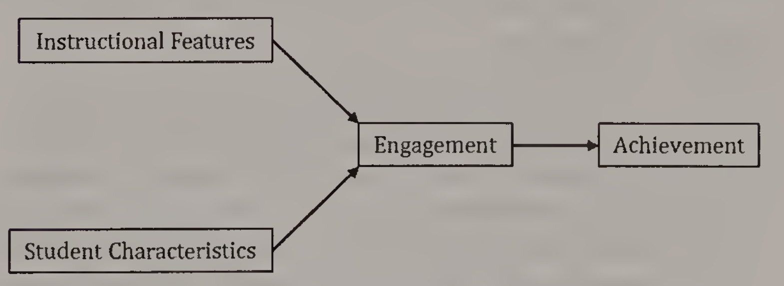


Figure 1. A model of academic learning.

for learning (Berliner, 1984; Karweit, 1984). Within engaged learning time, furthermore, researchers have come to focus on *productive learning time*—time in which the learner is exerting effort to learn on an appropriate academic task (Berliner, 1984). For example, van Gog, Ericsson, Rikers, and Paas (2005) point to the role of *deliberate practice*—spending extended periods of time effortfully engaged in tasks at an appropriate level of challenge that allow for continual improvement. Early work by Anderson (1993) provides an exemplary demonstration of the role of practice time in learning with computer-based cognitive tutors, and current work continues to demonstrate the positive impact of solving practice problems in e-learning (Clark & Mayer, 2011).

Although learning mechanisms were not highlighted in the early conceptions of time-on-task, the updated version of the time-on-task hypothesis is consistent with the idea that meaningful learning requires active cognitive processing in working memory during learning such as attending to relevant information, mentally organizing it into a coherent structure, and relating it to relevant prior knowledge activated from long-term memory (Mayer, 2011). What turns learning time into productive learning time is that the learner is engaged in appropriate cognitive processing on appropriate tasks during learning—processing that leads to constructing new knowledge and skills.

Based on these revisions in the classic concept of time-on-task, we expand the time-on-task hypothesis to focus also on the quality of time spent on homework. Overall, we examine three predictions about the relation between homework and achievement concerning the quantity of time (i.e., how much) and the quality of time (i.e., when).

1. How much: The most straightforward prediction of the time-on-task hypothesis is that time spent solving homework problems is related to course grade. However, a problem with traditional research on homework is that some studies use self-reported estimates of time spent doing homework. An important improvement in the current technology-supported study is that we have access to the actual time that students were working on their homework problems, including when they started and ended each session.

2. When: In addition to focusing solely on time spent on homework, a more sophisticated approach is to measure the quality of the time, such as the degree to which the homework activity was performed in advance of the deadline for submission. Although traditional research on homework generally does not include measures of when the homework was done, our technology-supported environment allows us to test the prediction that doing homework farther in advance of the deadline is related to course grade.

3. How many: In addition to focusing solely on time spent on homework, a more sophisticated approach is to measure how the time was spent. This challenge is problematic with traditional research on homework that does not involve in-process measures of homework activity. However, in our technology-supported environment, a straightforward way to measure the amount of effort put into doing homework is to count the number of strokes performed in solving homework problems. This allows us to test a more focused version of the time-on-task hypothesis: number of strokes performed while solving homework problems is related to course grade.

We examine these three predictions, and related predictions, across three cohorts of engineering students enrolled in an introductory course in statics.

Related Research on Data Mining in Education

Educational data mining with computer-based instructional systems has a rich history dating back to large-scale studies of computer-assisted instruction (CAI) in schools in the 1960s (e.g., Atkinson, 1968), extensive use of log files for modeling student learning with computer-based cognitive tutors (Anderson, 1993), and the subsequent use of log files with intelligent tutoring systems (Koedinger, D'Mello, McLaughlin, Pardos, & Rosé, 2015). In recent years, researchers have made significant progress in educational data mining or EDM (Koedinger, D'Mello, McLaughlin, Pardos, & Rosé, 2015; Romero, Romero, Luna, & Ventura, 2010). Much of the data used in this work is extracted from log files of intelligent tutoring systems (Beal & Cohen, 2008; Li, Cohen, Koedinger, & Matsuda, 2011; Mostow, González-Brenes, & Tan, 2011; Shanabrook, Cooper, Woolf, & Arroyo, 2010; Stevens, Johnson, & Soller, 2005; Trivedi, Pardos, Sráközy, & Heffernan, 2011) and learning management systems such as Moodle and Blackboard (Krüger, Merceron, & Wolf, 2010; Romero, Ventura, Vasilyeva, & Pechenizkiy, 2010). This work relies on a variety of data mining techniques including clustering (Antonenko, Toy, & Niederhauser, 2012; Stevens et al., 2005; Trivedi et al., 2011), model prediction (Li et al., 2011; Mostow et al., 2011; Stevens et al., 2005), and sequence analysis (Beal & Cohen, 2008; Kruger et al., 2010; Romero, Romero, et al., 2010; Shanabrook et al., 2010).

Our work differs from this in that we record and mine data from learning activities involving writing on paper, rather than activities involving typing on a computer keyboard. The work of Oviatt, Arthur, and Cohen (2006) suggests that natural work environments are critical to student performance. In their examinations of computer interfaces for completing geometry problems, they found that "as interfaces departed more from familiar work practice . . . , students would experience greater cognitive load such that performance would deteriorate in speed, attentional focus, metacognitive control, correctness of problem solutions, and memory" (p. 191). Similarly, Anthony, Yang, and Koedinger (2008) found that handwriting interfaces were more beneficial than keyboard interfaces for math tutoring systems. Mueller and Oppenheimer (2014) made a similar finding in relation to note-taking. They examined student note-taking using both longhand and laptops, and found that the latter can lead to shallower processing. Lectures were shown on a screen, with students taking notes, followed by distractor tasks. Using a model including both word count and verbatim overlap (three-word chunks from student notes matching the lecture transcript), they were able to predict performance on a test of the lecture material with a correlation coefficient of $r = .41$.

Macfadyen and Dawson (2010) mined data from a learning management system (LMS) to predict final course grade. Their best model was able to explain 33% of the variance in grade

utilizing three features: the number of mail messages sent, the number of assessments finished, and the total number of discussion messages posted. This provides some insights about the relationship between studying and course performance. However, the type of data available from a LMS—such as records of downloading course materials and submitting electronic assignments—does not provide a direct measurement of students' homework activity. We use smartpens to capture a fine-grained record of students' handwritten homework.

Researchers have used video recording to analyze students' problem-solving activities (Blanc, 1999; Hall, 2000). While this approach provides a detailed record of student work, the analysis is time-consuming. For example, Blanc (1999) made 75 recordings of students solving mathematics problems, but analyzed only two of the recordings. This sort of video analysis would be intractable in our studies, which involve hundreds of students completing homework throughout a quarter-long course. For our studies, smartpens provide a convenient and scalable approach for capturing high-resolution, timestamped records of problem-solving work.

There have been prior studies examining learning activities in statics. For example, work by Steif and Dollár (2009) examined usage patterns of a web-based statics tutoring system and found that learning gains increased with the number of tutorial elements completed. Similarly, work by Steif, Lobue, Kara, and Fay (2010) examined whether students can be induced to talk about the bodies in a statics problem, and if doing so can increase a student's performance. They used tablet PCs to record the students' spoken explanations and their handwritten solutions, but the written work was left mostly unanalyzed.

Researchers have only recently begun using smartpens for assessment. For example, Herold and Stahovich (2012) used smartpens to examine the homework of students who were asked to provide self-explanations for their solutions to statics problems. The study found that students who generated self-explanations were more likely to complete homework problems in the order assigned (i.e., complete one problem before beginning the next) than were students who did not generate self-explanations.

Our work builds on that of Rawson and Stahovich (2013) who used smartpens as part of a technique for making early predictions of student success or failure in a statics course. They used smartpens to record students' work on one homework assignment and a corresponding quiz given early in the course. They computed a number of features from this digital ink data including, for example, the total time spent on the homework and the amount of ink written. By themselves, these features were only weakly predictive of a student's course performance. However, when combined with a concept inventory score (Steif & Dantzler, 2005), these features produced useful early predictions.

In our work, we employ many of the ink features they developed. However, our goals are different. While their goal was to use data collected at the beginning of a course to make early predictions of success and failure, ours is to understand the relationship between homework habits and course performance. Our analysis considers homework behavior over the entire duration of a course, while they considered work from only a single assignment and quiz.

Recently, Van Arsdale and Stahovich (2012) demonstrated that the spatial and temporal organization of a student's solution to an engineering problem is indicative of the correctness of that solution. They recorded students' work on exam problems using smartpens and characterized the problem-solving activity in terms of the sequence of problem-solving steps and the arrangement of the work on the page. While they focused on a microscale analysis of problem-solving behavior on individual exam problems, we consider a macroscale analysis of homework habits over the duration of a course.

Herold, Stahovich, and Rawson (2013) used smartpens to examine the correlation between effort on a homework assignment and grade on that assignment. They characterized effort in terms of the amount of time spent and the amount of ink written. They also examined transfer from homework problems to subsequent homework, quiz, and exam problems. They characterized problem-solving work by the amount of time the pen was in contact with the paper, which is only a fraction of the time spent on the problem. They found that this "writing time" was correlated with performance on subsequent problems. Our work is similar in that we also examine the relationship between homework activity and success. However, we consider a longer time scale and our focus is understanding how homework habits over an entire course relate to success in that course.

Method

Participants and Course Setting

The participants were three cohorts of undergraduate engineering students at the University of California, Riverside who were enrolled in an entry-level course in statics—92 students in the winter quarter of 2010 (Year 1), 109 students in the winter quarter of 2011 (Year 2), and 127 students in the winter quarter of 2012 (Year 3). The winter term is the first offering of the statics course for the academic year. The majority of the students in the course are from mechanical engineering, although students from several other engineering majors, including materials science and environmental engineering, also take the course. Mechanical engineering students typically take the course in the sophomore year. The course includes two 80-min lecture periods per week. Students also attend a 50-min discussion section each week. The course employs a traditional lecture format.

Statics is the part of engineering mechanics focused on the equilibrium of objects subject to forces. The solution to a statics problem typically includes free body diagrams and equilibrium equations. The former represent the forces acting on a system, while the latter are the application of Newton's Second Law. Figure 2 shows a typical homework problem from the course and Figure 3 shows the sort of solution a student might generate for this problem. This image was constructed from digitized pen strokes captured with a smartpen.

In Year 1 students used Newton's Pen, an intelligent tutoring system for statics (Lee, Stahovich, & Calfee, 2011). This system was utilized during several discussion periods.

In Year 2 there were four separate discussion sections, each of which was provided with one of three different experimental treatments. Students from two discussion sections were asked to provide self-explanations for the problem-solving steps for six of the homework assignments. These students were provided with self-explanation prompts for these assignments. Students in a third

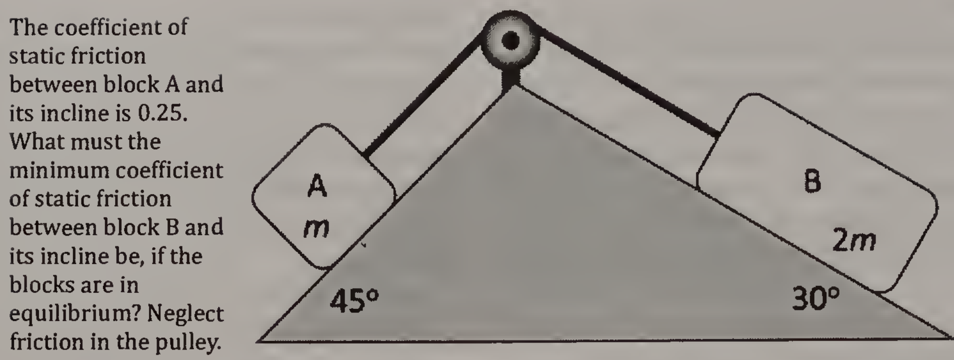


Figure 2. A typical statics problem.

discussion section used Newton's Pen during some of the discussion periods. The fourth discussion section served as the control. Students in this section did not provide self-explanation, nor did they use Newton's Pen.

In Year 3 students were randomly assigned to one of six experimental groups. Four of the groups were asked to provide self-explanation for the problem-solving steps on their homework. Each of these four groups was provided with varying amounts of scaffolding for self-explanation. Students in a fifth group used Newton's Pen during some discussion periods. Students in the sixth group served as the control. For the final homework assignment, all students were prompted to provide self-explanation without scaffolding. Also, in some discussion periods, students were given problems to solve. They began the problems in discussion, and if necessary completed them later. They submitted these solutions with their homework.

Course grade for all three cohorts was based on the following weighting: 10% for the homework score, 10% for the quiz score, 10% for the project score, 20% for the first midterm exam score, 20% for the second midterm exam score, and 30% for the final exam score. The exams and quizzes were not identical across cohorts but the content and format were similar. For example, for all 3 years the first midterm included one problem requiring students to compute a moment, one problem involving equilibrium analysis of a two-dimensional system, and one problem involving equilibrium of a three-dimensional system. All problems, except an ethics problem on the final exam, required free-form solutions, which typically required one or more free body diagrams and equilibrium equations. Problems were graded using a rubric that examined the correctness of the major elements of the solution. For example, an equilibrium problem might include a free body diagram, geometric calculations, and equilibrium equations. The

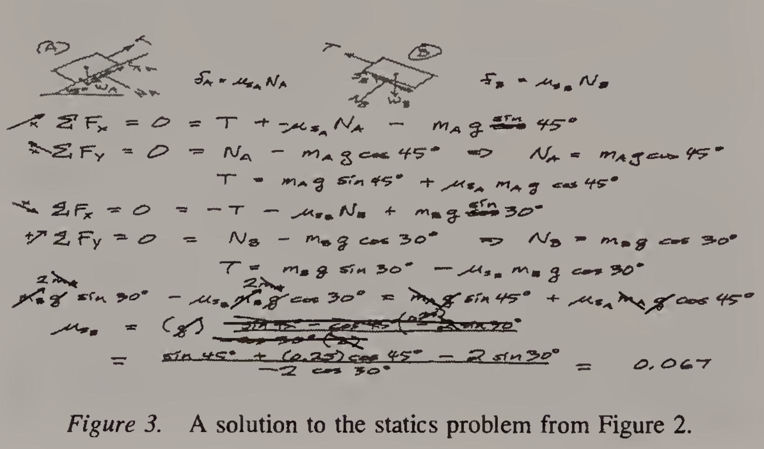


Figure 3. A solution to the statics problem from Figure 2.

credit for the problem would be divided over these elements according to their complexity, with more points being assigned to the more challenging elements. If an element was missing, the student would receive no credit for that element. Points were deduced from each element for various types of errors such as sign errors, missing terms (e.g., a missing force in a force equilibrium equation), incorrect terms (e.g., using "sine" instead of "cosine"), and so forth.

Procedure

Beginning in the third week of the course, students used smartpens to complete all homework assignments, quizzes, and exams. Students were instructed to use their smartpen instead of a pencil. We did not collect data from the first two homework assignments and quizzes. In Years 1 and 2, there were a total of nine homework assignments, and we collected data from the last seven. In Year 3, there was a total of eight assignments and we collected data from the last six. In all years we collected data from five quizzes (all quizzes except the first two), two midterm exams, and one comprehensive final exam. In Year 1, the seven homework assignments comprised a total of 41 problems, in Year 2 there were 44 problems, and in Year 3 there were 40. The instructor was aware of the general goal of the study—to capture student problem-solving data from the homework that could be related to course performance—but the data were not analyzed until after each cohort completed the course and received their final grades, thus eliminating the possibility of bias in assigning grades.

Livescribe Smartpens create two records: ink on paper and timestamped digitized pen strokes. In Year 1, students submitted both the paper copy of each assignment and their smartpens. We extracted the data from the smartpens and returned them to the students so they could complete their next assignment. For Year 2, we developed software to enable students to submit their assignments electronically. To do this, a student docked the smartpen to a PC using a USB cable. Our software then extracted the ink data and submitted it to a server for grading. We graded the homework electronically and returned it as a PDF. In Year 2, electronic submission was optional. Students could still submit the paper copy of an assignment, in which case we extracted the ink data form the smartpen at the end of the quarter. To encourage students to submit their work electronically, for some assignments the due date for electronic submission was several hours later than for paper submission. In Year 3, all students were required to submit their work electronically. However, if a student had technical

difficulties submitting a particular assignment, he or she could still submit it on paper and we extracted the ink data at the end of the quarter.

In Years 2 and 3, some students provided self-explanation with their homework. As self-explanation was not the focus of this project, and to maintain consistency across all students, we excluded the ink for the explanations from our analysis. However, we did include the self-explanation ink from the last assignment in Year 3, as all students provided self-explanation for that assignment. In Year 3, some homework submissions included problems that were solved in part during a discussion period. We excluded these problems from our analysis as they are not typical homework problems.

The Livescribe Smartpens have two clocks. One is used to display the current time of day, while the other is used to create timestamps for the pen strokes. The former can be adjusted, while the latter cannot. Having a nonadjustable clock for time stamps ensures that the time of the pen strokes is correct, even when there is a change to or from daylight saving time, for example. We used the time of an exam to determine the offset between the timestamp clock and the actual time of day. For Year 3, we also directly measured the offset before distributing the pens to the students. With this calibration approach, the offset of the timestamp clock is accurate to within about 5 min, which is adequate for our purposes.

For all 3 years, we conducted a survey at the end of the course with questions about demographics, study habits, and perceptions about the course and instructional technology used.

Data Mining With Smartpen Technology

We developed software to enable us to manually partition students' ink data into the individual problems comprising each assignment, quiz, and exam. The software renders the ink data on a computer display, enabling one to navigate through the pages of writing. A mouse is used to select the ink for an individual problem and assign a problem number to it. We then use software (Lin, Stahovich, & Herold, 2012) to automatically label each pen stroke as either an equation, free body diagram, or cross-out stroke (see Figure 3).

Once the digital ink has been partitioned into problems and labeled, we computed 13 quantitative measures to characterize a student's homework activity, as summarized in Table 1. Our first measure, *total homework time*, is the total time spent to complete all of the homework assignments. We define the time to complete one assignment as the time from the first pen stroke of the assignment to the last, excluding any periods of inactivity longer than 10 min. Any long inactivity periods partition the homework effort into *sessions*. Consecutive pen strokes within a session are never more than 10 min apart, while strokes from different sessions are always at least 10 min apart.

We use three measures to characterize the time effort over the assignment period. *Due date ink fraction*, computed as the fraction of the pen strokes written within 24 hr of the due date, measures the extent to which students wait until the "last minute" to complete an assignment. Similarly, *late night ink fraction*, computed as the fraction of the pen strokes written between midnight and 4 a.m., measures the fraction of work done late at night. Finally, *number of homework sessions* is simply the total number of sessions required to complete the assignments, with a new session counted when there is at least a 10-min break from the previous pen stroke.

In addition to considering the amount of time spent on homework, we also consider the amount of writing. As the name suggests, *total strokes* is the total number of pen strokes written to complete the assignments. We also count the number of *equation strokes*, the number of *diagram strokes*, and the number of *cross-out strokes*. These measures are computed using the auto-labeler from Lin, Stahovich, and Herold (2012). In addition to stroke count, we also consider the length of the pen strokes. *Total ink length*, which is computed in units of inches, is the total distance the pen tip travels on the paper.

We use three measures to characterize effort on individual homework problems. *Problems attempted* is the number of problems for which the student wrote at least 50 pen strokes. It is unlikely that a student made significant progress on a problem if he or she wrote fewer strokes than this. For example, simply writing "Problem 1" takes at least eight strokes. *Average time per problem* is the ratio of *total homework time and problems attempted*. This provides a means of comparing the effort of students even if they did not complete the same number of problems. *Average pen speed* is the ratio of *total ink length and total homework time*. This measure characterizes the pace of the work. Finally, the *out of order* measure describes the frequency with which a student works nonsequentially. Prior work has found

Table 1
Thirteen Measures Derived Through Smartpen Technology

Measure	Description
Total homework time	Total time to complete all assignments
Due date ink fraction	Proportion of pen strokes written within 24 hr of due date
Late night ink fraction	Proportion of pen strokes written between midnight and 4 a.m.
Number of homework sessions	Number of sessions used to complete the assignments
Total strokes	Number of pen strokes written to complete the assignments
Equation strokes	Number equation pen strokes written to complete the assignments
Diagram strokes	Number of diagram pen strokes written to complete the assignments
Cross-out strokes	Number of cross-out pen strokes written to complete the assignments
Total ink length	Total distance (in inches) the pen travels on paper for all assignments
Problems attempted	Number of problems for which the student wrote at least 50 pen strokes
Average time per problem	Total homework time divided by problems attempted
Average pen speed	Total ink length divided by total homework time
Out of order	Number of times a student transitions to a problem other than the next one in the assignment

that expert students often solve problems in the order assigned, while novice students may begin one problem and then move on to another before completing the former (Herold, Stahovich, Lin, & Calfee, 2011). The *out of order* measure is the number of times a student transitions to a problem other than the next one. For example, the sequence of problems 1, 3, 1, 2 has an *out of order* value of two. The transitions from 1 to 3 and 3 to 1 are nonsequential.

When computing these measures, we exclude any ink that was written more than 5 min prior to the time the homework assignment was posted. This tolerance compensates for the 5-min uncertainty in our timestamp clock calibration. We also exclude any ink written more than an hour after an assignment due date. As our electronic submission system did not prevent late submissions, some students did submit their homework late. We include any pen strokes written during this past-due hour in the *due date ink fraction*.

One of the questions on the end-of-class survey asked students to report the amount of time it took on average to complete a homework assignment, which we used to compute self-reported time on homework. For the first 2 years, the available choices for answering the question were: less than 2 hr, 2–4 hr, 4–6 hr, 6–8 hr, 8–10 hr, and more than 10 hr. In Year 3, the choices were reduced by one so that the last choice was "more than 8 hr." When computing the total time spent on homework, we consider a student's average assignment time to be the midpoint of the selected interval. However, if the student selected the largest choice, we use the lower bound (i.e., either 8 or 10 hr). For example, if a student in Year 1 reported "2–4 hr," we would compute the total self-reported time over the seven homework assignments to be 21 hr. Similarly, if they reported "more than 10 hr" we would compute the value to be 70 hr.

Results

Data Set

Our dataset includes data on 13 measures from a total of 328 students: 92 from Year 1, 109 from Year 2, and 127 from Year 3. All of these students completed the course and received a final course grade. We excluded data from one student in Year 2 and four from Year 3 because their digital ink data was corrupted.

As described in the Method section, some students in Years 2 and 3 were asked to write self-explanations and some others used an intelligent tutoring system. We wanted to determine whether the same pattern of results could be obtained in different contexts. We performed one-way analysis of variance (ANOVA) to determine if these treatments led to any significant differences in final grades between the experimental and control groups. In both cases, the differences were not significant ($p = .706$ for Year 2 and $p = .957$ for Year 3) and thus, in our analysis, we ignore these distinctions between students.

Table 2 shows the correlation between each of the 13 measures and course grade for all students, and for each cohort separately, with significant correlations at $p < .05$ denoted with an asterisk. We focus on the results for all students, and view the cohort data as a form of replication. Table 3 shows the means and standard deviations of each of the 13 measures for all students, and for each cohort separately.

Some of our measures are sensitive to the number of problems assigned. As the number of homework problems varied between the three cohorts, we performed another analysis in which we normalized the features by the number of problems assigned to the cohort. Four features—*due date ink fraction*, *late night ink fraction*, *average time per problem*, and *average pen speed*—did not require normalizing as they are insensitive to the number of problems. Normalizing the measures produced only a negligible change in the correlation with course grade. The correlations changed by less than .01 (and p by less than .003) for all measures.

We also investigated whether gender is significant to course performance. For Cohort 1 the average score for male students was .71 ($n = 78$), while the average score for female students was .65 ($n = 12$). However, this difference in means was nonsignificant, with $p = .212$. Similarly for Cohort 2, the average score for male students was .66 ($n = 87$), while the average score for female students was .63 ($n = 16$). This difference in means was again nonsignificant, with $p = .530$. For Cohort 3, the average score for male students was .68 ($n = 105$), while the average score for females was .61 ($n = 19$). This difference between means was significant, with $p = .028$.

Table 2
Correlation Between Course Grade and Each of 13 Smartpen Measures for all Students and Each Cohort Separately

Measure	All students	Cohort 1	Cohort 2	Cohort 3
Total homework time	.44*	.42*	.59*	.31*
Due date ink fraction	−.32*	−.38*	−.48*	−.20*
Late night ink fraction	−.06	−.08	−.15	−.04
Number of homework sessions	.33*	.05	.58*	.27*
Total strokes	.49*	.55*	.60*	.40*
Equation strokes	.49*	.54*	.61*	.40*
Diagram strokes	.41*	.46*	.51*	.34*
Cross-out strokes	.32*	.33*	.34*	.33*
Total ink length	.42*	.44*	.50*	.39*
Problems attempted	.45*	.35*	.68*	.27*
Average time per problem	.33*	.32*	.39*	.29*
Average pen speed	−.02	.02	−.11	.07
Out of order	.10	−.17	.27*	.09

* $p < .05$.

Table 3
Means and Standard Deviations for Each of 13 Smartpen Measures for all Students and Each Cohort Separately

Measure	All students		Cohort 1		Cohort 2		Cohort 3	
	μ	σ	μ	σ	μ	σ	μ	σ
Total homework time (hr)	17.1	8.5	17.7	6.4	16.9	9.1	17.0	9.2
Due date ink fraction	.7	.3	.7	.2	.7	.2	.6	.3
Late night ink fraction	.1	.1	.2	.2	.1	.1	.1	.1
Number of homework sessions	36.0	19.6	38.8	16.7	32.0	18.1	37.3	22.3
Total strokes	20189.5	9186.6	18944.0	6880.2	21162.0	10002.3	20257.0	9855.3
Equation strokes	14858.8	6939.1	13975.2	5411.3	15411.0	7330.4	15025.1	7542.9
Diagram strokes	4925.9	2373.6	4572.3	1708.0	5318.8	2767.8	4844.9	2390.9
Cross-out strokes	404.7	278.7	396.5	207.6	432.2	359.9	387.0	241.8
Total ink length (inches)	5936.2	3030.1	5568.8	2308.0	5979.0	3252.0	6165.7	3280.8
Problems attempted	34.4	8.9	35.7	5.9	35.5	10.2	32.4	9.2
Average time per problem (min)	29.1	11.3	29.5	9.8	27.6	10.8	30.0	12.6
Average pen speed (inches/second)	.104	.045	.093	.037	.105	.038	.112	.053
Out of order	20.2	15.3	17.7	11.1	22.6	19.6	20.1	13.5

We also examined the correlation between measures of prior knowledge and course performance. Here we use two measures to quantify prior knowledge: the student's SAT score (based on combined verbal, quantitative, and writing scores) and their high school GPA. For Cohort 1 (r = .534, $p < .001$) and Cohort 3 (r = .284, p = .003), there was a significant correlation between SAT score and final course grade, but not for Cohort 2 (r = .091, p = .378). The correlation between high school GPA and final course grade was significant for Cohort 1 (r = .317, p = .003) and Cohort 2 (r = .285, p = .004) but not for Cohort 3 (r = .184, p = .052).

We performed a stepwise linear regression to examine the predictive ability of our entire set of measures. In computing a stepwise model we required the probability of $F \leq .05$ to enter a measure, and the probability of $F \geq .10$ to remove a measure. We initialized the model by including three measures: *total strokes, total homework time*, and *problems attempted*. For all students, *total strokes, problems attempted, out of order* and *due date ink fraction* were selected with r = .57, $p < .001$. For Cohort 1, *total strokes* and *out of order* were selected with r = .67, $p < .001$. For Cohort 2, *total strokes, problems attempted*, and *due date ink fraction* were selected with r = .72, $p < .001$. For Cohort 3, only *total strokes* was selected with r = .40, $p < .001$. Thus, in the analysis with the best statistical power (i.e., the combined data from all students), there is evidence that each of four smartpen measures (i.e., *total strokes, problems attempted, out of order*, and *due date ink fraction*) makes a unique contribution to predicting course grade.

As a follow-up we conducted another stepwise linear regression identical to the one described previously, but with SAT score entered as the first variable and the smartpen variables entered in order of their correlation. For all students, *SAT score, text strokes, problems attempted, out of order* and *due date ink fraction* were selected with r = .63, $p < .001$. For Cohort 1, *SAT score, total strokes, out of order* and *late night ink fraction* were selected with r = .76, $p < .001$. For Cohort 2, *SAT score, problems attempted*, and *due date ink fraction* were selected with r = .71, $p < .001$. For Cohort 3, *SAT score* and *text strokes* were selected with r = .63, $p < .001$. Thus, in the analysis with the best statistical power (i.e., the combined data from all students), there is evidence that each of four smartpen measures (i.e., *text strokes, problems attempted, out of order*, and *due date ink fraction*) contribute uniquely to predicting course grade, even when the effects of prior knowledge are controlled (i.e., smartpen variables predict course grade beyond the effects of SAT score). Overall, although construction of a factor-analyzed measurement instrument based on smartpen variables is beyond the scope of this study, there are indications that course grade is uniquely predicted by a collection of smartpen measures.

The final course grade includes homework score with a weight of 10%. To examine if this artificially increased the correlations between our measures of homework activity and course grade, we recomputed the course grade excluding homework score and recomputed the correlations. This resulted in only a negligible change in the correlations, and no change in the factors chosen in the regression analyses. More specifically, for Cohorts 1 and 2, the changes in correlations were no greater than .04. For Cohort 3, the changes were no greater than .07.

How Much: Is Homework Time Related to Course Grade?

According to the basic version of the time-on-task hypothesis, students who spend more time working on their homework should get better grades in the course. The first line of Table 2 shows the correlation between total time spent on the homework problems and course grade for all students combined, and for each cohort separately. As the table illustrates, there is a significant correlation for each cohort and for all students combined, consistent with predictions. Overall, there is strong and consistent support for the time-on-task hypothesis, based on data collected through smartpen technology.

What happens when we look at students' self-reported time on homework per week as reported on a postquestionnaire? In contrast to the significant correlation between course grade and the actual time on homework recorded through smartpens, the correlation between course grade and self-reported time on homework is not positively significant for all students combined (r = −.16) nor for each of the three cohorts (r = −.29 for Cohort 1, r = −.14 for Cohort 2, and r = −.13 for Cohort 3). Instead, the correlation is negative for all three cohorts, and the negative correlation for Cohort 1 is statistically significant.

Furthermore, is the students' self-reported time consistent with the actual measured time to complete homework assignments? For two of the three cohorts, there was only a weak correlation between self-reported time and *total homework time*: For Cohort 1, $r = .21, p = .052$; for Cohort 2, $r = .16, p = .139$; and for Cohort 3, $r = .35, p < .001$. Additionally, nearly all students overreported their homework time. For Cohort 1, 88.5% of students overreported homework time with an average overestimation of 19.0 hr. For Cohort 2, 85.5% of students overreported homework time with an average overestimation of 23.5 hr. For Cohort 3, 85.5% of students overreported homework time with an average overestimation of 13.4 hr.

This pattern of differences between actual time and self-reported time points to the value of technology-supported measures of homework activity in testing the time-on-task hypothesis. This set of contrasting findings constitutes a major contribution of this study.

When: Is the Timeliness of Homework Activity Related to Course Grade?

According to the updated version of the time-on-task hypothesis, which considers the quality of the time spent on homework, students who commonly wait until the last minute to do homework (i.e., within 24 hr of the due date) or who commonly do homework late at night (i.e., midnight to 4 a.m.) should get worse grades in the course. Consistent with this prediction, the second line of Table 2 shows a significant negative correlation between *due date ink fraction* and course grade for all students, and for each cohort. In contrast, the third line of Table 2 does not show a significant correlation between *late night fraction* and course grade for any of the cohorts, suggesting perhaps that working late at night is not necessarily an indication of lower quality time. Overall, a major empirical contribution is strong and consistent evidence that the quality of how homework time is spent (as measured by the proportion of homework time done within 24 hr of the deadline) is related to course grade. The smartpen technology allows us to address this prediction of an updated version of the time-on-task hypothesis.

Similarly, breaking an assignment up into multiple sessions may be a way to enable distributed practice—spreading practice over multiple sessions—which has been shown to improve learning (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). Accordingly, time-on-task should be most efficient when it is spread over multiple sessions. Consistent with this prediction, the fourth line of Table 2 shows that *number of homework sessions* correlates significantly with course grade for all students combined and for two of the three cohorts. Again, smartpen technology allows us to address a prediction about the time-course of homework activity using data that is not otherwise available.

How Many: Is the Amount of Writing Activity Related to Course Grade?

According to the updated version of the time-on-task hypothesis, which considers the amount of productive activity, students who create more pen strokes while working on homework assignments should get better grades in the course. Consistent with this prediction, the fifth line of Table 2 shows a significant correlation between *total strokes* and course grade for all students, and for each cohort. Also consistent with predictions, the next lines in Table 2 show the same pattern of significant correlations (for all cohorts) between grades and *equation strokes, diagram strokes, cross-out strokes*, and *total ink length*, respectively. Overall, there is consistent evidence that higher achievement is related to the level of effort exerted by students as indicated by their pen strokes.

Similarly, Table 2 shows that for all students and for each cohort separately, there is a significant correlation between the *number of problems attempted* and course grade (line 10) and between *average time per problem* and course grade (line 11).

What is not related? Two variables did not correlate consistently with course grade—*average pen speed* and *out of order*—perhaps because they are not appropriate measures of the amount of productive activity. Writing faster or slower does not necessarily indicate more or less effort, and trying problems out of order can be attributed to several causes other than effort, including lack of concentration.

Discussion

Empirical Implications

Concerning issues about time-on-task, actual time spent working on homework problems was positively correlated with course grades, but self-reported time spent on homework problems was not. Concerning the time course of homework activity, the amount of homework time spent within 24 hr of the deadline was negatively correlated with course grades, the amount of homework time spent between midnight and 4 a.m. was not correlated with course grades, and breaking homework time into more sessions was positively correlated with course grades. Concerning actual behavior and effort on homework problems, course grades were positively correlated with the total number of pen strokes, equation strokes, diagram strokes, cross-out strokes, total ink produced, total problems attempted, and time per problem. Course grades were not consistently correlated with average pen speed or solving problems out of order.

Theoretical Implications

This study investigates a crucial link in a model of academic learning, the link between engagement or effort to learn, as measured by the amount and time students allocate to a learning task, and performance, as measured by learning outcome in a college course. In particular, the present study examines the idea that the amount of time that students spend in productive learning is related to academic achievement in a course. Although no causal conclusions can be drawn, the work draws attention to a potential causal mechanism leading to learning—namely, amount of productive learning activity. Importantly, both the quantity of time spent on homework and the quality of how homework time is used are related to achievement. Higher quality use of time is reflected in doing homework long before it is due and breaking assignments into smaller sessions. Effortful activity on homework is reflected in the number of pen strokes, the total ink produced, and the number of problems attempted. A major contribution of this project is to enable more detailed measures of student effort or engagement—which is proposed to be the mechanism underlying academic learning.

Practical Implications

This is a correlational study that examines actual performance in a real college course, so no causal conclusions can be drawn. However, this study offers preliminary evidence for the potential of homework as an aid to student achievement, particularly when students work on their homework in a timely and effortful way. The role of productive time on task has long been recognized as a critical issue in intelligent tutoring systems (Anderson, Corbett, Koedinger, & Pelletier, 1995).

Methodological Implications

This study highlights the potential of educational data mining techniques in general (i.e., techniques for measuring and summarizing learner activity during learning), and smartpen technology in particular, for educational research. Smartpen technology allows for assessing student study activity at a level of detail that was not previously possible, and thereby offers a new and powerful methodology for testing implications of educational theories.

Our results confirm what other researchers have proposed (Blumner & Richards, 1997; Schuman et al., 1985): Students' self-reports of study effort are often unreliable. Finally, this study points to the useful role of replication in educational research (as noted by Shavelson & Towne, 2002) by showing the same pattern of results across three independent cohorts of students.

Limitations and Future Directions

This work is a first step at building techniques that can provide automated assessment of performance from an analysis of handwritten homework. Our present analysis examines the relationship between the amount of effort on homework and performance. In future work, we plan to examine how patterns of homework activity contribute to success. For example, our current analysis suggests that doing homework just before it is due may not be a successful strategy, but causal claims cannot be drawn based on the correlational relationship identified in this study. Thus experimental work is needed to test causal claims. Experimental research should be designed to explicitly test hypotheses suggested by this study concerning the possible causal role of productive time on task by directly manipulating this factor and examining the effects on learning outcome. Within the context of experimental research, future work is also needed to examine potential moderating variables such as the learner's prior knowledge and metacognitive skills.

The correlations involving self-reported homework time and actual homework time should be interpreted in light of the fact that our self-report measure of homework time involved asking the learner to check a category that was converted to a number for analyses (such as "two to four hours per week" being recorded as 3, "four to six hours per week" being recorded as 5, etc.), whereas our smartpen measure of homework time is based on a continuous scale.

Another concern is whether the act of being asked to use smartpens, and the ensuing awareness of being observed, could cause students to be more careful about how they deal with homework problems than they would otherwise, could make them want to use scratchpaper before solving homework problems with a smartpen, and could create discomfort or distraction that affect homework behavior. In short, it is important to ensure that students do their homework in their usual way but with use of their smartpens and nothing else. In the present study, students were instructed to show all their work using their smartpen, and a postexperimental questionnaire indicated reasonable compliance. On a survey from Cohort 3 asking students to rate smartpen use on a scale of 1 ("doing all homework elsewhere") and 7 ("using the pen to do everything") the mean rating was 5.1 (SD = 1.7). Future work should involve more evidence concerning fidelity, such as poststudy interviews. Similarly, the total time measure was not based on any activity before the first pen stroke so it would not include time to initially read and think about the problem before starting to answer. Another thorny issue concerns whether course grade is an adequate measure of learning outcome. In the present study, course grade was based on tests that involved concepts related to the homework problems, but no detailed method of alignment was implemented.

The present work is based on the idea that a deeper analysis of the sequencing of homework activities can provide additional insights about successful and unsuccessful study strategies. Identifying such strategies will lead to experimental studies that ultimately enable automated coaching systems to examine students' study habits and recommend interventions aimed at increasing academic success. Additionally, we plan to extend the smartpen technology to the study of note-taking during classroom lectures in order to identify classroom learning strategies that are related to course grade.

This work is a step in applying educational data mining techniques to learning activities in traditional, rather than online, environments. Our current studies have focused on one course (i.e., statics), and more work is needed to determine how our techniques will generalize to other domains for which homework assignments comprise handwritten problem solving. We anticipate that our techniques will be applicable to assessing homework habits in a variety of math, science, and engineering subjects.

References

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences, 4,* 167–207. http://dx.doi.org/10.1207/s15327809jls0402_2

Anthony, L., Yang, J., & Koedinger, K. R. (2008). Toward next-generation intelligent tutors: Adding natural handwriting input. *IEEE MultiMedia, 15,* 64–68. http://dx.doi.org/10.1109/MMUL.2008.73

Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development, 60,* 383–398. http://dx.doi.org/10.1007/s11423-012-9235-8

Atkinson, R. C. (1968). Computerized instruction and the learning process. *American Psychologist, 23,* 225–239. http://dx.doi.org/10.1037/h0020791

Beal, C. R., & Cohen, P. R. (2008). Temporal data mining for educational applications. *Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence, 5351,* 66–77.

Berliner, D. (1984). The half-full glass: A review of research on teaching. In P. L. Hosford (Ed.), *Using what we know about teaching* (pp. 51–84). Alexandria, VA: Association for Supervision and Curriculum Development.

Berliner, D. (1991). What's all this fuss about instructional time? In M. Ben-Peretz & R. Bromme (Eds.), *The nature of time in schools*. New York, NY: Teachers College Press.

Blanc, P. (1999). Solving a non-routine problem: What helps, what hinders. *Proceedings of the British Society for Research into Learning Mathematics, 19,* 1–6.

Blumner, H. N., & Richards, H. C. (1997). Study habits and academic achievement of engineering students. *The Journal of Engineering Education, 86,* 125–132. http://dx.doi.org/10.1002/j.2168-9830.1997.tb00275.x

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64,* 723–733.

Clark, R. C., & Mayer, R. E. (2011). *e-Learning and the science of instruction* (3rd ed.). San Francisco, CA: Pfeiffer.

Cooper, H. (1989). *Homework*. White Plains, NY: Longman. http://dx.doi.org/10.1037/11578-000

Cooper, H. (2001). *The battle over homework* (2nd ed.). Thousand Oaks, CA: Corwin.

Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research 1987–2003. *Review of Educational Research, 76,* 1–62. http://dx.doi.org/10.3102/00346543076001001

Credé, M., & Kuncel, N. R. (2008). Study habits, skills, and attitudes: The third pillar supporting collegiate academic performance. *Perspectives on Psychological Science, 3,* 425–453. http://dx.doi.org/10.1111/j.1745-6924.2008.00089.x

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4–58. http://dx.doi.org/10.1177/1529100612453266

Ebbinghaus, H. (1964). *Memory*. New York, NY: Dover. (Original work published 1885)

Graham, S., & Williams, C. (2009). An attributional approach to motivation in school. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 11–34). New York, NY: Routledge.

Hall, R. (2000). Video recording as theory. In D. Lesh & A. Kelley (Eds.), *Handbook of research design in mathematics and science education* (pp. 647–664). Mahwah, NJ: Lawrence Erlbaum.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge.

Herold, J., & Stahovich, T. (2012). Characterizing students handwritten self-explanations. *Proceedings of the American Society for Engineering Education 119th Annual Conference and Exposition*, San Antonio, TX.

Herold, J., Stahovich, T., Lin, H.-L., & Calfee, R. C. (2011, June). The effectiveness of "pencasts" as an instructional medium. *Proceedings of the American Society for Engineering Education 118th Annual Conference and Exposition*, Vancouver, British Columbia, Canada.

Herold, J., Stahovich, T. F., & Rawson, K. (2013). Using educational data mining to identify correlations between homework effort and performance. *Proceedings of the American Society for Engineering Education 120th Annual Conference and Exposition*, Atlanta, GA.

Jones, L., & Ruch, G. M. (1928). The twenty-seventh yearbook of the national society for the study of education: Nature and nurture, part ii-their influence upon achievement. In G. M. Whipple (Ed.), *Achievement as affected by amount of time spent in study* (pp. 131–134). Bloomington, IL: Public School Publishing Co.

Karweit, N. (1984). Timed-on-task reconsidered: Synthesis of research on time and learning. *Educational Leadership, 41,* 32–35.

Keith, T. Z. (1982). Time spent on homework and high school grades: A large-scale path analysis. *Journal of Educational Psychology, 74,* 248–253. http://dx.doi.org/10.1037/0022-0663.74.2.248

Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., & Rosé, C. P. (2015). Data mining and education. *Wiley Interdisciplinary Reviews: Cognitive Science, 6,* 333–353. http://dx.doi.org/10.1002/wcs.1350

Krüger, A., Merceron, A., & Wolf, B. (2010). A data model to ease analysis and mining of educational data. In R. S. J. d. Baker, A. Merceron, & P. I. Pavlik (Eds.), *Proceedings of the 3rd International Conference on Educational Data Mining* (pp. 131–140). Pittsburgh, PA.

Lee, C.-K., Stahovich, T., & Calfee, R. C. (2011). A pen-based statics tutoring system. *Proceedings of the American Society for Engineering Education 118th Annual Conference and Exposition*, Vancouver, British Columbia, Canada.

Li, N., Cohen, W. W., Koedinger, K. R., & Matsuda, N. (2011). A machine learning approach for automatic student model discovery. In M. Pechenizkiy, T. Calders, C. Conati, Š. Ventura, C. Romero, & J. Stamper (Eds.), *Proceedings of the 4th International Conference on Educational Data Mining* (pp. 31–40). Eindhoven, the Netherlands.

Lin, H., Stahovich, T. F., & Herold, J. (2012). Automatic handwritten statics solution classification and its applications in predicting student performance. *Proceedings of the 119th American Society for Engineering Education Annual Conference and Exposition*, San Antonio, TX.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education, 54,* 588–599. http://dx.doi.org/10.1016/j.compedu.2009.09.008

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511811678

Mayer, R. E. (2011). *Applying the science of learning*. Upper Saddle River, NJ: Pearson.

Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 43–71). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139547369.005

Mostow, J., González-Brenes, J. P., & Tan, B. H. L. (2011). Learning classifiers from a relational database of tutor logs. In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, & J. Stamper (Eds.), *Proceedings of the 4th International Conference on Educational Data Mining* (pp. 149–158). Eindhoven, Netherlands.

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science, 25,* 1159–1168. http://dx.doi.org/10.1177/0956797614524581

Oviatt, S., Arthur, A., & Cohen, J. (2006). Quiet interfaces that help students think. *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology - UIST '06* (pp. 191–200). New York, NY: ACM Press.

Rawson, K., & Stahovich, T. F. (2013, June). Predicting course performance from homework habits. *Proceedings of the American Society for Engineering Education 120th Annual Conference and Exposition*, Atlanta, GA.

Romero, C., Romero, J., Luna, J., & Ventura, S. (2010, June 11-13). Mining rare association rules from e-learning data. In R. S. J. d. Baker, A. Merceron, & P. I. Pavlik Jr. (Eds.), *Proceedings of the 3rd International Conference on Educational Data Mining* (pp. 171–180). Pittsburgh, PA.

Romero, C., Ventura, S., Vasilyeva, E., & Pechenizkiy, M. (2010). Class association rules mining from students' test data. In R. S. J. d. Baker, A. Merceron, & P. I. Pavlik Jr. (Eds.), *Proceedings of the 3rd International Conference on Educational Data Mining* (pp. 131–140). Pittsburgh, PA.

Schuman, H., Walsh, E., Olson, C., & Etheridge, B. (1985). Effort and reward: The assumption that college grades are affected by quantity of study. *Social Forces, 63,* 945–966. http://dx.doi.org/10.1093/sf/63.4.945

Schunk, D. H., & Pajares, F. (2009). Self-efficacy theory. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 35–54). New York, NY: Routledge.

Shanabrook, D. H., Cooper, D. G., Woolf, B. P., & Arroyo, I. (2010). Identifying high-level student behavior using sequence-based motif discovery. In R. S. J. d. Baker, A. Merceron, & P. I. Pavlik Jr. (Eds.), *Proceedings of the 3rd International Conference on Educational Data Mining* (pp. 191–200). Pittsburgh, PA.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education*. Washington, DC: National Academy Press.

Steif, P. S., & Dantzler, J. A. (2005). A Statics Concept Inventory: Development and psychometric analysis. *The Journal of Engineering Education, 94,* 363–371. http://dx.doi.org/10.1002/j.2168-9830.2005.tb00864.x

Steif, P. S., & Dollár, A. (2009). Study of usage patterns and learning gains in a web-based interactive static course. *The Journal of Engineering Education, 98,* 321–333. http://dx.doi.org/10.1002/j.2168-9830.2009.tb01030.x

Steif, P. S., Lobue, J. M., Kara, L. B., & Fay, A. L. (2010). Improving problem solving performance by inducing talk about salient problem features. *The Journal of Engineering Education, 99,* 135–142. http://dx.doi.org/10.1002/j.2168-9830.2010.tb01050.x

Stevens, R., Johnson, D. F., & Soller, A. (2005). Probabilities and predictions: Modeling the development of scientific problem-solving skills. *Cell Biology Education, 4,* 42–57. http://dx.doi.org/10.1187/cbe.04-03-0036

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4419-8126-4

Trivedi, S., Pardos, Z. A., Sárközy, G. N., & Heffernan, N. T. (2011). Spectral clustering in educational data mining. In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, & J. Stamper (Eds.), *Proceedings of the 4th International Conference on Educational Data Mining* (pp. 129–138). Eindhoven, Netherlands.

Van Arsdale, T., & Stahovich, T. (2012, June). Does neatness count? What the organization of student work says about understanding. *Proceedings of the American Society for Engineering Education 119th Annual Conference and Exposition*, San Antonio, TX.

van Gog, T. (2013). Time on task. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 432–433). New York, NY: Routledge.

van Gog, T., Ericsson, K. A., Rikers, R. M. J. P., & Paas, F. (2005). Instructional design for advanced learners: Establishing connections between the theoretical frameworks of cognitive load and deliberate practice. *Educational Technology Research and Development, 53,* 73–81. http://dx.doi.org/10.1007/BF02504799

Xu, J. (2013). Homework and academic achievement. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 199–201). New York, NY: Routledge.

Received September 28, 2015
Revision received April 5, 2016
Accepted April 8, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 2, 220–232

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000127

Benefits of Guided Self-Management of Attention on Learning Accounting

Seedwell T. M. Sithole and Paul Chandler
University of Wollongong

Indra Abeysekera
CQ University

Fred Paas
Erasmus University Rotterdam and University of Wollongong

This research investigated the effects of 3 instructional design formats on learning introductory accounting. In accordance with cognitive load theory, it was predicted that students who would learn with a guided self-managed instructional design format would outperform students who would learn with a conventional split-attention format or an integrated format on a recall test and a transfer test. In the guided self-management condition students were instructed to reorganize text and diagrams to reduce the need to search the solution steps within the text and match them with corresponding parts of the diagram, thereby freeing cognitive resources for learning. The results of an experiment conducted with 123 undergraduate university students confirmed the hypothesis by consistently demonstrating that students in the guided self-managed condition outperformed students in the integrated and split-attention conditions on the recall and transfer tests.

Keywords: accounting, cognitive load theory, learning, guided self-management, split-attention

Cognitive load theory (CLT; Ayres & Sweller, 2005; Paas, Renkl, & Sweller, 2003; Sweller, 2015; Sweller, Ayres, & Kalyuga, 2011) uses knowledge of human cognition to provide instructional design principles that support the efficient use of working memory. Over the last three decades CLT research has almost exclusively focused on instructor-managed cognitive load, on how instructional designers can best design learning materials following CLT principles (Paas, Van Gog, & Sweller, 2010). The basic idea is that when the instructional principles are used by instructional designers they lead to decreased working memory load caused by task aspects and activities that are unproductive for learning, thereby freeing up working memory resources for activities that are productive for learning.

The study reported in this article takes a different approach and follows up on recent research focusing on equipping learners with strategies to self-manage cognitive load when dealing with instructional materials with evident split attention (see also Agostinho, Tindall-Ford, & Roodenrys, 2013; Roodenrys, Agostinho, Roodenrys, & Chandler, 2012). Despite the empirically convincing superiority of integrated formats over split-source formats (Ginns, 2006), recent studies of Agostinho and colleagues (2013) and Roodenrys and colleagues (2012) have suggested that learners can also be trained to manage their own cognitive load when confronted with split-attention materials.

The split-attention effects is one of the major design principles of CLT, which indicates that replacing multiple sources of visual information with a single, integrated source of information leads to better learning (Ayres & Sweller, 2005; Ginns, 2006). Separated sources of information such as text and diagram require the learner to hold small segments of text in memory while searching for the matching diagram. Such a process continues until all the information is rendered intelligible (Agostinho et al., 2013).

Models and Learning From Multiple Representations

Learning environments frequently combine several forms of representations like animations, pictures, texts, tables, or formulas. However, learners' acquisition of knowledge from these multiple sources requires learners to create referential connections between corresponding elements and corresponding structures in the different representations to construct a coherent mental representation (e.g., Schwartz & Martin, 2004; Seufert, 2003). Many studies investigating learning from multiple representations use multimodal theories of human memory to formulate hypotheses and explain results (see, e.g., Liu, Lin, Tsai, & Paas, 2012). Two influential models that have inspired many other theories, such as cognitive load theory (Plass, Moreno, & Brünken, 2010; Sweller, 2015) and the cognitive theory of multimedia learning (Mayer, 2009), are Baddeley's working memory model (Baddeley, 1992) and Paivio's dual-coding model (Clark & Paivio, 1991; Paivio, 1986). Baddeley's model divides working memory into a "visual-spatial scratch pad" for dealing with visually based information and a "phonological loop" to deal with auditory, primarily speech-based, information. These two systems, in turn, are governed by a

This article was published Online First April 28, 2016.

Seedwell T. M. Sithole, Department of Accounting, School of Accounting, Economics and Finance, University of Wollongong; Paul Chandler, Early Start Research Institute, University of Wollongong; Indra Abeysekera, Department of Accounting, School of Business and Law, CQ University; Fred Paas, Department of Psychology, Erasmus University Rotterdam and Early Start Research Institute, University of Wollongong.

Correspondence concerning this article should be addressed to Seedwell T. M. Sithole, Department of Accounting, School of Accounting, Economics and Finance, University of Wollongong, Building 40, Northfields Avenue, Wollongong NSW 2522, Australia. E-mail: seedwell@uow.edu.au

central executive. In Paivio's model, pictorial information and verbal information processed in different cognitive subsystems: An imagery system and a verbal system. Pictures are processed and encoded both in the imagery system and in the verbal system, while words and sentences are generally processed and encoded only in the verbal system. The memory-enhancing effect of pictures in texts is attributed to the benefit of a dual coding as compared with single coding in memory.

Despite the benefits associated with learning from multiple representations such as elaboration, flexibility and multiple perspectives, there is empirical evidence suggesting that learners, particularly those with little prior knowledge have difficulties to integrate multiple sources of information and typically fail to construct a coherent mental representation (Mayer, 1997; Rau, 2015; Stern, Aprea, & Ebner, 2003). In this study we use cognitive load theory to explain this failure and to provide a potential solution in terms of guided self-management of cognitive load.

CLT and Its Instructional Formats

CLT is based on the assumption that human cognitive architecture consists of a working memory with very limited capacity when dealing with new information (Sweller, 2015) and an unlimited long-term memory in which elements are organized and stored in the form of domain-specific knowledge structures known as schemas (van Merriënboer & Ayres, 2005). Although working memory can only hold a limited number of elements at any given time, working memory can access complex schemas consisting of huge arrays of interrelated elements allowing a learner access to previously learned material stored in the long-term memory (Sweller, 2015). Access of schemas in long-term memory can reduce working memory load, thereby freeing working memory resources for learning. However, learning content that is novel to students does not have associated schemas in long-term memory.

CLT suggests that effective instructional design should take into account human cognitive architecture, in particular concentrating on effective use of limited working memory resources (Choi, Van Merrienboer, & Paas, 2014; Sweller, 2015). One of the loads identified by CLT is extraneous cognitive load, which is the burden imposed on working memory by the manner in which the information is presented or the activities in which the learner must engage (Sweller et al., 2011). This load can result from poorly designed instructional material. CLT research has mainly focused on instructor-manipulated instructional materials and providing instructor designed learning materials that take into account the cognitive load imposed on the learner.

When learning new content, instructional formats having a separate text and diagram hinder learning because they require a learner to search relevant text and match it with particular sections of the diagram. Such a presentation unnecessarily overloads the limited capacity of working memory that is not relevant to the learning process (Leahy, Chandler, & Sweller, 2003). Hence many studies have illustrated the importance of instructional material designed with CLT principles in mind. Five of the most researched CLT derived instructional effects are the (a) expertise reversal effect (e.g., Blayney, Kalyuga, & Sweller, 2010; Kalyuga, Ayres, Chandler, & Sweller, 2003). (b) worked example effect (e.g., Paas & Van Gog, 2006; Sweller, 2006), (c) split-attention effect (e.g., Ayres & Sweller, 2005; Clark, Ngyuen, & Sweller, 2006), (d) modality effect (e.g., Ginns, 2005; Goolkasian, Foos, & Eaton, 2009), and (e) redundancy effect (e.g., Samur, 2012). For an exhaustive overview of CLT-based instructional formats and their empirical base, see Sweller et al. (2011) and van Merriënboer and Sweller (2005). These design principles have been verified in numerous experiments conducted with a diverse range of instructional materials. Within the cognitive load theory framework, one main characteristic that has been identified as a contributor to the negative effect on learning is the need for learners to split-attention between multiple sources of information that must be integrated before they can be understood (Ayres & Sweller, 2005; Clark et al., 2006).

To summarize, empirical research has provided valuable insights into different facets of learning, for example, demonstrating that the process of coherence formation is cognitively demanding and learners with insufficient prior knowledge are often unable to cope with this task. Consequently, they do not use different representations but rather concentrate only on one representation and therefore fail to integrate and reach the goals of elaboration, abstraction, flexibility, and coherence (Seufert, 2003; Seufert & Brünken, 2006).

Previous research has focused on how educators can use CLT designed instructional material to manage students' cognitive load and improve their learning performance, but the current research investigates student application of CLT design principles to manage their own cognitive load and improve learning. Research on learning from split-attention instructional materials has not extensively provided techniques that would empower students to successfully and systematically deal with split-attention materials on their own. This study investigates whether it is possible to instruct novice students on how to self-manage the split attention. With an exclusive focus on novice students using accounting instructional material, two cognitive load theory effects that may result from manipulating those materials, that is, the split-attention effect and guided self-management effect were explored. The next sections discuss the split attention effect in the context of accounting instructional material, followed by a discussion on guided self-management instructional designs. Thereafter, an experiment is presented for the evaluation of the guided self-management strategy.

The Split-Attention Effect

Separate text and diagrams are very difficult to understand and consequently have a negative effect on learning (Ginns, 2006). This form of presentation, referred to as split-attention, demands the learner's effort to mentally reorganize the text or related explanations related to the diagram and/or symbols. Split-attention occurs when texts accompanying a diagram are presented separately and are unintelligible in isolation. To understand the material, the learner must hold small pieces of text in working memory while searching for the matching relevant diagrammatic representation. This process continues until all the information is rendered intelligible.

Over the last two decades researchers have been developing alternatives to instructional formats that require learners to extensively search and match that increases the load in working memory (Ayres & Sweller, 2005; Florax & Ploetzner, 2010; Morrison, Dorn, & Guzdial, 2014). One successful strategy to reduce work-

ing memory load imposed by search and match activities that are not relevant to learning is physical integration of the different information sources (Ayres & Sweller, 2005; Paas et al., 2003; Sweller, 2015; Sweller et al., 2011). In a diagram and text presentation, the separation of the text from the diagram forces the learner to look back and forth between the relevant parts of the diagram and the text, a process that unnecessarily imposes working memory load. If the diagram is combined with text, the learner could concentrate better on learning the content from the combined presentation (Sweller et al., 2011). Research has shown that integrated formats are superior to split-attention formats even though learners are required to mentally integrate under both formats (Agostinho et al., 2013; Paas et al., 2003; Sweller, 2015).

A meta-analysis of the split-attention effect has shown that integrated instructional formats reduce extraneous cognitive load (Ginns, 2006). Replacing multiple sources of information with a single, integrated source of information assists with more effective learning (Ginns, 2006; Mayer, 2009). The split-attention effect (e.g., Rose & Wolfe, 2000) can arise because information is spatially (spatial contiguity effect; e.g., Clark & Mayer, 2008) and temporally separated (temporal contiguity effect; e.g., Ginns, 2006). These studies have shown that students often learn more when complex educational content is designed to reduce the space (spatial contiguity effect) or time (temporal contiguity effect) between disparate but related elements of learning content.

An example of integration of instructional material is when fragments of text are directly embedded into a diagrammatic presentation or as close as possible to corresponding components of a diagram (see Figure 1). As illustrated in Figure 1, the first diagram (i), presents an example of a conventional split-attention worked example in geometry. In Figure 1, the diagram that is above the text outlines the solution to the problem. The diagram and text are presented separately. In processing the information on the diagram and text below it, the learner has to understand the solution steps and then link them with the diagram. This requires mentally integrating the two sources of information, drawing on considerable cognitive resources from the learner, contributing to learning difficulty.

In the integrated worked example (see Figure 1, second diagram [ii]), the learner can allocate working memory resources to the relational dimensions of the problem, because his or her mental capacity is released from the need to search and match the solution steps and link them with the diagram. The integrated example enhances learning since it guides the learner through the steps of a worked example (Ayres & Sweller, 2005; Ginns, 2006; Mayer, 2009).

While a lot of instructional materials make use of both a diagrammatic component and a textual component of information, which imposes a high demand on working memory, the current most efficient method of dealing with split-attention is physically integrating instructions (Agostinho et al., 2013). This manner of presentation represents a form of instructor-manipulated intervention (Paas et al., 2010). The argument proposed in this article is that guided self-management, in which learners are asked to link text to relevant parts in the diagram may be an alternative to instructor integrated instructional materials. Particularly in the current educational environment where learners can access vast amounts of information, it is likely that learners will often be confronted with split-attention learning materials without any form of instructional guidance. In those cases learners will have to self-manage learning from the split-attention materials.

Self-Management Effect

The self-management effect was recently developed within a cognitive load framework by Roodenrys and colleagues (2012; see also Agostinho et al., 2013). The researchers involved noticed the high variability of instructional formats used on the World Wide Web and elsewhere and the high likelihood of these materials being designed without any cognitive load considerations. Largely, two options are available to cognitive load theorists. The first is to keep on reconstructing deficient instructional formats (e.g., split attention) into more effective formats (e.g., integrated). However, given the sheer amount of information that now exists electronically, most of it generated without cognitive considerations, there

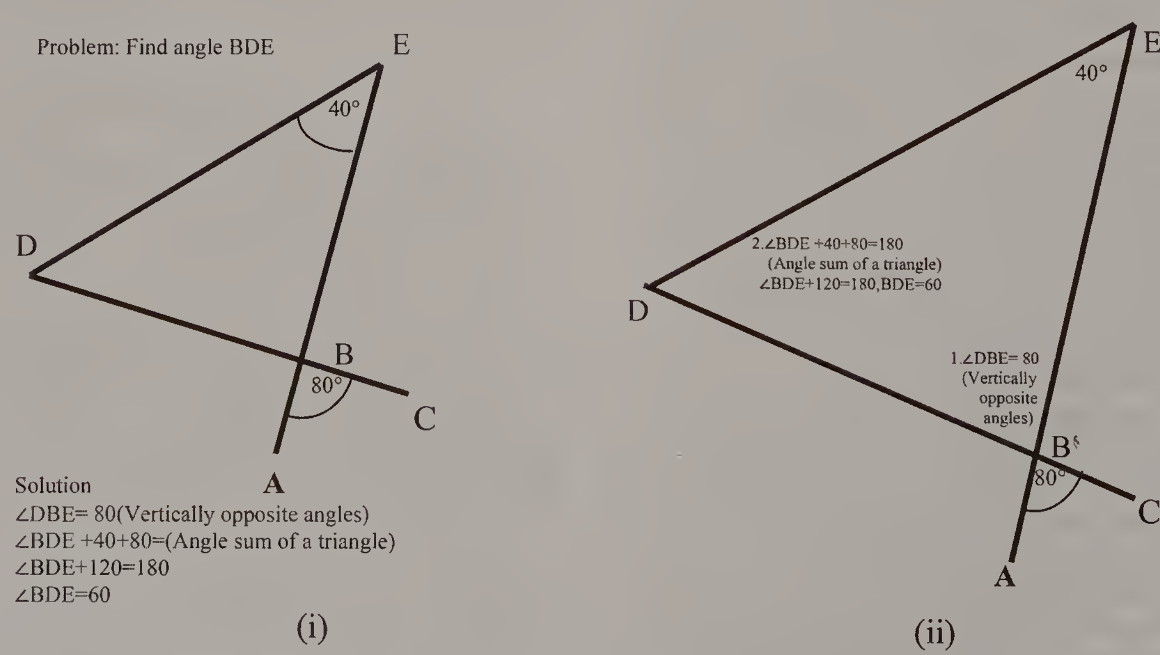


Figure 1. Split attention format (i) and Integrated format (ii). *Source:* Ayres and Sweller (2005, p. 208).

are now severe limitations that instructors can realistically achieve to facilitate learning.

Thus, the field has started to move from simply reconstructing deficient formats to look at areas of self-regulated learning as a means of controlling learners' own cognitive load Although there is an extensive literature on active integration performed by learners (see, e.g., Aleven, McLaren, Roll, & Koedinger, 2006; Berthold & Renkl, 2009; Bjork & Bjork, 2011; Bodemer, Ploetzner, Feuerlein, & Spada, 2004; Mason, Caterina, & Pluchino, 2013), the guided-self-management strategy that is investigated in this study specifically focuses on self-management of cognitive load and is considered a general skill that can be transferred by learners to other domains. Guided self-management is a very specific and direct form of instructional guidance designed to facilitate the skill of self-management of learning. Learners can be given very specific coaching into identifying inefficient instructional formats and then given examples of how to self-manage such load (Agostinho, Tindall-Ford, & Bokosmaty, 2014; Gordon, Tindall-Ford, Agostinho, & Paas, 2016; Roodenrys et al., 2012; Tindall-Ford, Agostinho, Bokosmaty, Paas, & Chandler, 2015). The results of a number of studies have indicated that not only does this self-management work, in the long run it may also be more effective than any other cognitive load generated format for assisting transfer. The article will now discuss some of the research in the area of the self-management effect.

Self-management consists of making connections between the two representations (e.g., text and diagram) through annotations such as highlighting, underlining, and drawing arrows (Agostinho et al., 2014, 2013; Roodenrys et al., 2012). Most self-management studies involve the use of split-attention self-management techniques with either paper-based materials or in an online environment, drawing arrows, numbering or moving text to related diagrammatic components. Self-management in the experiment conducted in this article required the learner to use paper-based materials to highlight, underline, and draw arrows to parts of the diagram that the learner understands are related, the aim being to reduce learners' need to search and, thus, freeing cognitive resources for learning.

Confronting learners with conditions of learning that impose initial challenges to the learner (i.e., desirable difficulties) have been found to enhance retention and transfer performance (Bjork & Kroll, 2015). For example creating small discrepancies between an auditory narration and on-screen text can be desirable. In the study of Bjork and Kroll (2015) participants studied a lesson about the life cycle of a star that comprised animation, narration, and on-screen text. When the narration was a little different from the on-screen text, participants learned more than when the narration and text were identical (Bjork & Kroll, 2015; Yue, Bjork, & Bjork, 2013). Paas and Van Merriënboer (1994) also showed that students who studied high-variability worked examples invested less time and mental effort in practice, and attained better and less effort-demanding transfer performance than students who solved high-variability conventional problems.

Tindall-Ford et al. (2015) examined secondary school students self-managing the split-attention when learning about the properties of angles in mathematics. They found that the students who received instructor guidance to integrate text with a diagram performed better on later tests than students who received no such guidance. Using educational psychology materials, Roodenrys et al.'s (2012) experiments showed that it is possible to instruct students on how to self-manage information. In their first experiment, participants in an integrated group performed significantly better than the self-management group across recall and far transfer performance items. For near transfer items, the self-management group slightly outperformed the integrated group. Roodenrys et al. (2012) also concluded that self-management instructions need to be carefully constructed so that the instructions would not result in an unnecessary cognitive load because of either split-attention or redundancy. Their studies also showed that the positive effects of self-managed instructions may be demonstrated on transfer tasks.

Thus, to this point, the research within the self-management effect has provided some promising performance over the split-attention format in areas such as educational psychology (Roodenrys et al., 2012), mathematics (Tindall-Ford et al., 2015), and educational technology (Agostinho et al., 2013). The current study investigated whether a guided self-managed group in accounting would be superior to a split-attention group.

The Current Study

Based on the above discussion, the current study examined the hypothesized superior learning outcomes of guided self-management instructions and integrated instructions over conventional split-attention instructions. The hypotheses regarding the effects of instructional condition on performance and mental effort are described next.

Instructional Design and Performance

1. Performance by split-attention format group (Group 1) and guided self-managed group (Group 3):

Hypothesis 1a: Students in the guided self-managed format will outperform students in the split-attention format on recall tests.

Hypothesis 1b: Students in the guided self-managed format will outperform students in the split -attention format on transfer tests.

Within cognitive load research, experimental evidence has shown that novices' learning to solve problems by studying split-attention material leaves little processing capacity for schema acquisition and the capability to recall and transfer knowledge (e.g., Ayres & Sweller, 2005). Most recent research (Agostinho et al., 2014; Roodenrys et al., 2012; Tindall-Ford et al., 2015) has revealed that self-managing split-attention problems (such as in Group 3) results in superior performance as compared with learners who have not been provided with any guidance.

Instructional Design and Mental Effort

Subjective mental effort scores, which are considered a reliable measure of overall cognitive load (Paas et al., 2003), were collected in this study. Based on CLT and the empirical findings of previous studies into the self-management effect the following predictions regarding the effects of instructional design on mental effort were formulated.

The first hypothesis in relation to mental effort was tested in the first part of the experiment. Participants in Part 1 of the Experiment were given specific guidance to assist them with moving text as close as possible to the associated diagram. This is expected to lead to higher cognitive load compared with the other conditions. It was predicted that:

Hypothesis 2: Students in the guided self-managed format group (Group 3) will report higher effort (cognitive load) than students in the split-attention format group (Group 1), and the integrated format group (Group 2).

Cognitive load is increased by the need to mentally integrate various sources of information (Ayres & Sweller, 2005). Hence the split-attention format group should report higher cognitive load than the integrated format group. The guided self-management group is expected to report a high cognitive load because of the need to move the text as close as possible to the associated diagram.

In relation to the transfer task (Part 2 of the Experiment), participants in the guided self-managed format (Group 3) were expected to use the guidance learned in Part 1 of the Experiment to move text as close as possible to the associated diagram without specific guidance from the instructor. This is expected to lead to higher cognitive load compared with the other conditions. It was predicted that:

Hypothesis 3: Students in the guided self-managed format (Group 3) would use guidance to self-manage and report higher cognitive load than students in the integrated format group (Group 2) and split-attention group (Group 1).

Research has shown that learners have the capability of self-managing instructional materials and perform better on test items compared with the split-attention group (Agostinho et al., 2014; Roodenrys et al., 2012; Tindall-Ford et al., 2015). This result has come out despite the fact that the guided self-management group is required to carry out an additional task of moving text as close as possible to the diagram during the learning phase. The research conducted to date shows that learners who are taught to self-manage instructional materials on their own perform better than the split-attention format (Agostinho et al., 2013; Roodenrys et al., 2012). Learners who self-managed split attention performed the same as the integrated group (Roodenrys et al., 2012; Tindall-Ford et al., 2015). To test the above hypotheses, an experiment was conducted in the current study.

Method: Part 1 of the Experiment

The aim of the experiment was to inquire into the split-attention effect in accounting instructional materials. In addition, the experiment sought to test whether specific guidance developed to assist participants with moving text as close as possible to the associated diagram in split-attention learning materials could lead to higher learning performance than a traditional split-attention condition.

Participants and Design

The participants were 123 first-year undergraduate students (63 men and 60 women, M = 21 years old, SD = 2.17) from a Zimbabwean university. Approval for human subjects research was obtained from the Human Research Ethics Committee at the Zimbabwean university. A power analysis using the Gpower computer program (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that a total sample of 40 people would be needed to detect large effects (d = .8) with 97% power using a t test between means with α at .05. Participants were enrolled in 13 degree programs, each taking an introduction to accounting course. Consented participants were randomly assigned to one of the three conditions. There were 41 students in the split-attention group (Group 1; 22 men and 19 women, M = 22 years old, SD = 2.22), 40 students in the integrated group (Group 2; 21 men and 19 women, M = 20 years old, SD = 1.48), and 42 students in the guided self-management group (Group 3; 20 men and 22 women, M = 21 years old, SD = 2.80). Students participated voluntarily in the study, and were not paid for participation. They had been informed of the study 1 week before the experiment being conducted.

The experiment was conducted during Week 3 of a 13 week semester with students studying a first year accounting core subject. At the start, before the experiment commenced, the researchers explained the organization and reasons for the experiment. Students were informed that participation was voluntary and that the results from the experiment were not part of the subject's assessment, and that data would be collected anonymously. There was no incentive offered to the participants. Participants were given participant information sheets and consent forms. They signed the consent form stating their written agreement to take part in the study. Students who agreed to participate were randomly assigned to one of the three conditions.

Before responding to the test questions, a pretest questionnaire was distributed. The participants answered questions about their age, gender, first language, and knowledge of accounting. This took 10 min to complete.

The 123 participants in both the first and second study ranged in gender from 48% to 53% men, and from 47% to 52% women in each instructional format group. The dominant language spoken by the participants is Shona with over 93% in each instructional format. Participants' gender and linguistic homogeneity was thus apparent across the groups. All students had passed a high school formal English language examination. The students' language proficiency was sufficiently high to respond to questions in English. Each participant was then tested individually with researchers supervising the test. Part 1 of the Experiment took 45 min.

Materials and Procedure

The instructional materials explained the basic accounting equation, the debit and credit rules, and their effect on the basic accounting equation. The instructional materials were obtained from an accounting textbook (Weygandt et al., 2010, pp. 53–54) in the form of split-attention, but formatted as follows for each of the three conditions:

Group 1–split attention: The instructional material used by Group 1 (split-attention format) was similar to that found in the textbook. An example of the material used in the current study is illustrated in Figure 2.

Group 2–integrated group: The instructional material in Group 2 (the integrated format) was presented in a format that integrated the diagram with the text (see Figure 2). The

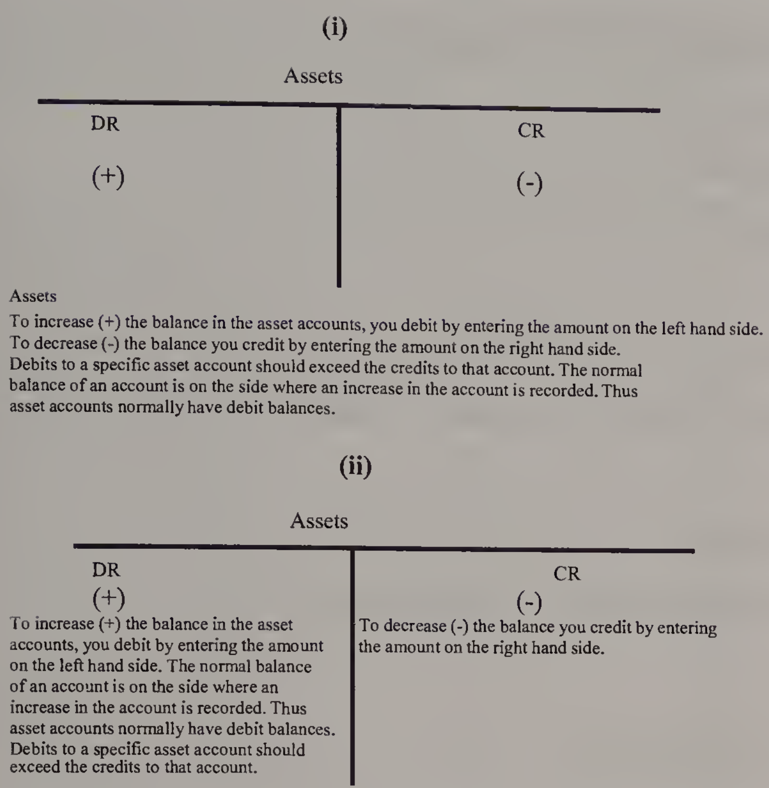


Figure 2. Example of conventional split-attention format (i) and integrated format (ii) in accounting.

content was reformatted to decrease split-attention by bringing the text as close as possible to the diagram (integrating). The integrated material was developed after reviewing the research concerning split-attention (e.g., Agostinho et al., 2013; Ayres & Sweller, 2005; Roodenrys et al., 2012; Tindall-Ford et al., 2015) and then reformatting the instructional material. An example of the material used in the current study is illustrated in Figure 2.

Group 3–self-managed cognitive load: Instructional materials in Group 3 (the guided self-managed format) were developed in such a way that it assisted participants to integrate the diagram with the text. An example of the material used in the current study is illustrated in Figure 3. The material contained guidance (As shown in Figure 4) such as: (a) Draw a circle around the information for each debit and credit; (b) Draw an arrow to link it to its corresponding place on the diagram. (c) Highlight with a highlighter, or underline, mark circles on key words, number with a pencil or pen in sequence on the diagram and on the text. Participants in Group 3 were explicitly asked to implement the guidance before attempting to learn the materials. The techniques for self-management were extensively researched by Roodenrys et al. (2012) and can be considered the common, current method using self-management of cognitive load.

The participants in Part 1 of the Experiment worked manually using pencil and paper. Part 1 of the Experiment had three phases: learning phase, test phase, and post phase. At the start of the experiment participants completed a pretest questionnaire. They received two A3 pages of learning materials that contained learning instructions. The learning instructions differed among the three groups. During the test, as they completed the test questions they responded to mental effort rating questions. The responses helped us to evaluate the extent of recall of learning content and transfer of knowledge by solving problems in different situations.

In the learning phase, the participants were given 15 min to review the learning materials provided to them. In the final phase, the researcher administered the test that was formatted as a single sided A4 booklet. The test consisted of 28 recall and 11 transfer items. The participants were given 45 min to complete the test.

An example of a recall question in the test phase is that students were asked to write the basic accounting equation. Recall questions required students to retrieve the acquired knowledge (Carpenter, 2012). An example of a transfer question is; In May, Company X records the transaction by a debit to Accounts Receivable for $5,000 and a credit to Service Revenues for $5,000. What is the effect of this entry upon the accounting equation for Company X? Tick the appropriate answer: Assets: Increase, Decrease, No Effect; Liabilities: Increase, Decrease, No Effect; Equity: Increase, Decrease, No Effect.

The transfer questions tested the ability to transfer acquired knowledge, and the demands of the questions were higher than recall questions. Transfer questions required a student to apply the

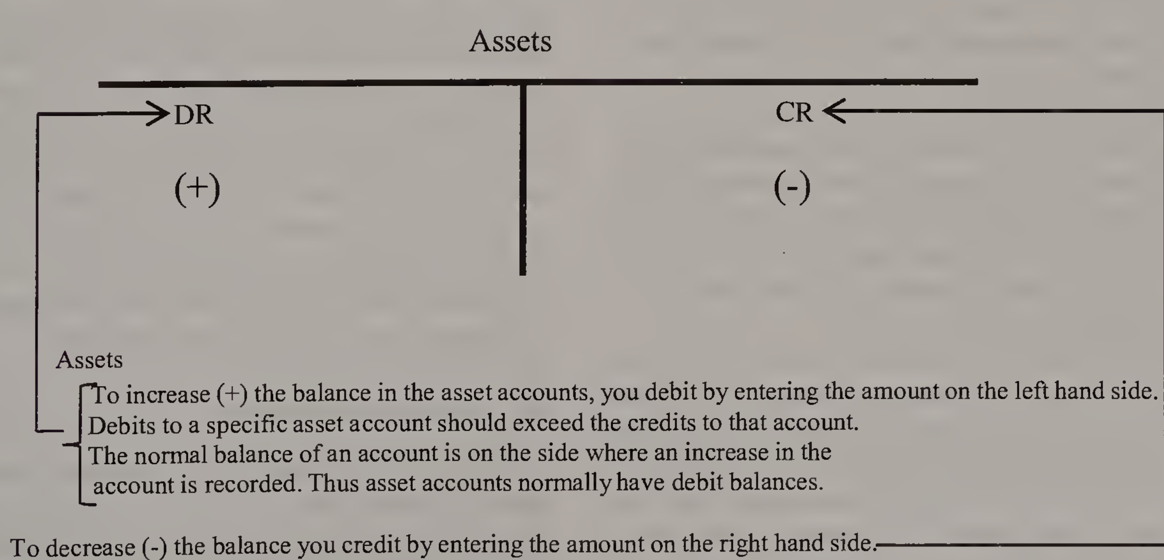


Figure 3. Example of self-management using arrows.

Self-management Guidance Experiment Part 1– Group 3

The specific steps below will assist you to learn the accounting equation more effectively by making use of your working memory.

Please complete the following tasks before you start reading the material presented:

(a) Draw a circle around the information for each debit and credit

(b) Draw an arrow to link it to its corresponding place on the diagram. An example has been done for you.

(c) Highlight with a highlighter, or underline, mark circles on key words, number with a pencil or pen in sequence on the diagram and on the text. An example has been done for you.

Figure 4. Guidance on self-management.

knowledge acquired during instruction to a novel situation (Mayer & Wittrock, 1996). Participants provided mental effort ratings after the learning phase and after attempting every question as outlined by Paas (1992). Participants wrote answers and any comments they wished to provide on the blank spaces immediately below the questions. The researcher collected all the test booklets soon after the students completed the tasks.

Pilot Study

A pilot study was conducted before the experiment. It aimed at refining instructional guidance, instructional content, and time that should be allowed for each phase of the studies. Five students from the same university took part. Those five students did not participate in the experiment. The time limit, for both the learning phase and test phase, was determined in the pilot study. The time given to complete the test was strictly controlled to avoid the possibility of a systematic difference in processing time between the split-attention, integrated and self-managed groups. Research has demonstrated that processing time is positively related to recall (Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007).

Rating of mental effort. After students completed the working through the instructional materials, they were asked to rate the cognitive load associated with the learning task. To measure perceived cognitive load, this study used Paas and Van Merrienboer's (1994) 9-point subjective cognitive load rating scale. This is an established scale to measure the level of overall cognitive load (Ayres & Paas, 2012; Van Gog & Paas, 2008).

Mental effort rankings were solicited from participants at the end of the learning phase and after each question in the test. For example, "How much mental effort did you invest to learn the material?" at the completion of the learning phase and 'How much mental effort did you invest to answer this question?' at the end of each test question. The ratings on the levels of mental effort of the accounting exercises were assumed to assess cognitive load indirectly (Paas et al., 2003; Van Gog & Paas, 2008).

Compliance measures. Compliance was an additional measure included in the analysis for participants allocated to Group 3 (the guided self-managed format) of Part 1 of the Experiment. Compliance refers to the participant's use of the guidance attached to the instructional materials. Evidence of compliance involved examination of the instructional materials (A3 sheets of paper) to determine if participants implemented the instructional guidance (to assist guided self-management). Participants were considered "compliant" if they highlighted material with a highlighter, underlined material, or marked circles on key words with a pencil or pen.

Reliability. In the present studies, the reliability of the scale was estimated with Cronbach's coefficient α. The data from the recall test scores and transfer test scores were entered and run in SPSS using the reliability analysis function. For the internal consistency check for the recall test scores and transfer test scores, all groups were first combined and then separated by treatment group to ensure that internal consistency for all groups was established. The experiment combined results displayed a high level of internal consistency. The results showed a recall Cronbach's α of .946 and .918 for the transfer task. The transfer Cronbach's α for the experiment was .892 and .882 for the transfer task. Overall, Cronbach's α for the guided self-management of cognitive load experiment ranged between .806 and .885 among the three treatment groups. The recall results ranged between .882 and .946 and the transfer Cronbach's α ranged between .806 and .892. Gall et al. (2003) states that for research purposes, having a reliability of .80 or higher is considered sufficiently reliable.

Results the Experiment

The data were analyzed with a one-way analyses of variance (ANOVAs) with code 1 (i.e., split-attention instruction), 2 (i.e., integrated instruction), and 3 (i.e., guided self-managed instruction) representing the levels of the between-subjects factor instructional format, to determine its effects on the dependent measures,

performance on recall, transfer, and perceived cognitive load. In case of significant F tests, pairwise post hoc comparisons using Tukey contrasts were conducted. The α level was set at .05 ($p <$ .05) when evaluating tests of statistical significance. To measure effect size, Cohen's d was calculated, with values of .10, .30, and .50 characterizing small, medium, and large effect sizes, respectively (Cohen, 1988).

Pretest response and analysis. A one-way ANOVA was conducted on pretest responses to explore differences across the three groups involved in the experiment. Means and *SD*s are shown in Table 1. The one-way ANOVA for pretest questions revealed no significant main effect of group for age; $F(2, 120) =$ .01, $p = 1$, and knowledge of accounting $F(2, 120) = 1.191$, $p =$ .307. The equivalence of knowledge of accounting between the groups was important to note, as no group came to the study with higher a priori knowledge of accounting. This is evidence that the three groups are equivalent on significant demographic dimensions. Accordingly, statistically significant differences detected later are more likely to be caused by differences between the treatment conditions.

Mental effort ratings. Means and *SD*s of mental effort ratings in the learning phase are shown in Table 1. Results from the one-way ANOVA for mental effort invested in the learning phase indicated significant differences across the three formats, $F(2, 120) = 75.77$, $p < .05$, effect size $\eta_p^2 = 0.55$. There were large and significant between-groups differences on mean mental effort rating in the learning phase, which indicated that the guided self-managed group reported lower levels of cognitive load than the integrated group. The perceived amount of mental effort invested in the split-attention format was higher than that invested with the integrated format.

Table 1
Means and SDs for Pretest Responses, Recall, and Transfer Test Scores as a Function of Instructional Condition

	Experiment			
	Part 1		Part 2	
Pretest	Mean	*SD*	Mean	*SD*
Age				
Split-attention ($n = 41$)	21.20	2.86		
Integrated ($n = 40$)	21.20	2.28		
Guided self-managed ($n = 42$)	21.19	2.79		
Knowledge of accounting[a]				
Split-attention ($n = 41$)	2.10	.37		
Integrated ($n = 40$)	1.98	.42		
Guided self-managed ($n = 42$)	2.02	.27		
Recall performance[b]				
Split-attention	49.93	13.32	51.46	16.84
Integrated	59.18	12.33	64.28	15.12
Guided self-managed	80.00	14.38	87.14	13.63
Transfer performance[c]				
Split-attention	36.73	16.38	40.71	17.69
Integrated	53.30	17.99	61.73	15.63
Guided self-managed	66.83	16.72	83.21	19.18
Mental effort rating				
Learning phase				
Split-attention	7.90	1.16	7.17	1.64
Integrated	6.80	1.77	4.73	2.04
Guided self-managed	4.02	1.46	3.62	1.91
Recall				
Split-attention	7.49	.98	6.56	1.66
Integrated	4.55	1.36	4.63	1.44
Guided self-managed	2.93	1.33	2.71	1.13
Transfer				
Split-attention	5.88	.93	6.05	1.38
Integrated	5.00	.94	3.89	1.3
Guided self-managed	3.52	.99	2.33	1.21

[a] Actual responses were 1 to 5 for knowledge of accounting, [b] actual raw score ranges were 0 to 28 for recall, and for [c] transfer it was 0 to 11.

A Tukey post hoc test for learning phase revealed that the mental effort factor was statistically significant, with the guided self-managed group recording the lowest cognitive load (4.02 rating, $p < .05$) compared with the integrated format group (6.80 rating, $p < .05$), $d =$ 1.71, and split-attention format group (7.90 rating, $p < .05$), $d = 2.94$. Tukey post hoc tests also revealed that the integrated format group reported a significantly lower level of perceived cognitive load compared with the split-attention format group, $d = 0.74$, indicating a large effect size.

Performance measures. Table 1 shows means and *SD*s for performance measures in the experiment based on one-way ANOVAs. A one-way ANOVA for recall scores showed a significant main effect for the recall test items; $F(2, 120) = 54.834$, $p < .05$, effect size $\eta_p^2 = 0.478$. Mean recall and transfer scores showed that the guided self-managed group had higher scores than the integrated group, which in turn had higher scores than the split-attention group. Consistent with predictions, post hoc comparisons using Tukey contrasts showed that the guided self-managed group performed significantly better than the split-attention group, $d =$ 2.17, and integrated group, $d = 1.45$ with the integrated group performing better than the split-attention group, $d = 0.73$, indicating a large effect size.

The one-way ANOVA for transfer questions also demonstrated a significant main effect of group; $F(2, 120) = 32.478$, $p < .05$, and effect size $\eta_p^2 = 0.351$. Post hoc comparisons using Tukey contrasts showed that the guided self-managed group performed significantly better than the split-attention group, $d = 1.82$, and the integrated group, $d = 0.78$. Again the integrated group performed better than the conventional split-attention group, $d = 0.96$.

Mental effort rating on instruction. After the learning phase and after each test question, students were asked to rank their effort in terms of perceived mental effort on recall and transfer questions. One-way ANOVAs were conducted to determine the influence of instructional methods on recall and transfer test performance.

Table 1 shows the mean ratings and *SD*s for the ratings in the test phase. There were large and significant differences arising from the different instructional formats based on mean values of recall results; $F(2, 120) = 144.973$, $p < .05$, effect size $\eta_p^2 =$ 0.707. Instructional format also differentially affected the mean values of transfer results; $F(2, 120) = 64.834$, $p < .05$, effect size $\eta_p^2 = 0.519$. The perceived cognitive load was significantly lower in the integrated group than in the split-attention group. Contrary to expectations, the guided self-managed group reported significantly lower levels of perceived cognitive load than the integrated group.

Follow-up Tukey post hoc tests for recall on the differences revealed that mental effort was statistically significant with the guided self-managed group recording the lowest cognitive load (2.93 rating, $p < .05$) compared with the integrated group (4.55 rating, $p < .05$), $d =$ 1,21 and split-attention group (7.49 rating, $p < .05$), $d = 3.90$. Tukey post hoc tests also revealed that the

integrated group (4.55 rating, $p < .05$) reported a significantly lower level of perceived cognitive load compared with the split-attention group (7.49 rating, $p < .05$), $d = 1.32$. Similarly a Tukey post hoc test for transfer items revealed that the cognitive load differed significantly between groups, with the guided self-managed group recording the lowest cognitive load (3.52 rating, $p < .05$) compared with the integrated group (5.00 rating, $p < .05$), $d = 1.53$, and the split-attention group (5.88 rating, $p < .05$), $d = 2.45$. A Tukey post hoc test also revealed that the integrated group (5.00 rating, $p < .05$) reported a significantly lower level of cognitive load than the conventional split-attention group (5.88 rating, $p < .05$), $d = 0.93$.

Guidance compliance. Results of the compliance measures indicated that 95% participants followed the guidance about how to self-manage split attention. Compliance referred to the participant's use of the guidance attached to the instructional materials for Group 3. Participants were considered "compliant" if they performed at least one of the tasks suggested, which were highlighting material with a highlighter, using arrows to link text and diagram, underlining material, or drawing circles to mark key words with a pencil or pen.

The most common strategy used by the participants was highlighting, underlining, or numbering (86%). The second most used strategy was drawing an arrow to link it to its corresponding place on the diagram. Only 36% of the participants drew a circle around the information. The use of at least these tasks is seemingly quite useful in understanding the instructional material. The level at which these tasks were conducted suggests that full utilization of the guidance contributed to higher performance scores.

Part 2 of the Experiment

The aim of the transfer task was to reinquire the existence of split-attention with new learning materials and test the transfer of guided self-management skills to a new learning domain. Part 2 of the experiment used the same procedure and the same participants as Part 1 to test the robustness of a possible "self- management effect." The instructional materials were changed, and learners were not specifically instructed anymore to follow a certain procedure.

Materials and Procedure

The instructional materials were about the topic of ratio analysis. In the test phase there were three pages of recall and transfer test questions to be answered, including a requirement to rate mental effort after answering every test question. The first set of instructional materials was taken directly from the textbook (Weygandt et al., 2010, pp. 783–785). This constituted the split-attention format. The second set of instructional materials was adjusted to reflect the integrated format and for the third set, the guided self-management format, no guidance was given. The students in the guided self-management format group were expected to utilize the guided self-management techniques they gained from Part 1 of the Experiment. Part 2 of the experiment proceeded directly after Part 1.

Rating of mental effort. Similar to Part 1 of the experiment, a 9-point subjective cognitive load rating scale was used.

Compliance measures. Compliance measures involved analysis of the participant's use of highlighting, underlining, or marking circles on key words with a pencil. Evidence of compliance involved examination of the instructional materials to determine if participants implemented the guidance.

Results of Part 2 of the Experiment

Mental effort ratings in the learning phase. Means and *SD*s of mental effort ratings for learning phase are shown in Table 1. Results from the one-way ANOVA for mental effort invested in the learning phase indicated significant differences across the three formats, $F(2, 120) = 39.04$, $p < .05$, effect size $\eta_p^2 = 0.394$. Consistent with predictions, there were large and significant between-groups differences on mean mental effort rating in the learning phase. Mean learning phase ratings showed that the guided self-managed group reported lower levels of cognitive load than the integrated group. The perceived amount of mental effort invested by the split-attention group was higher than that invested by the integrated group.

Tukey post hoc tests for the learning phase revealed that mental effort differed significantly between the groups, with the guided self-managed group recording the lowest cognitive load (3.62 rating, $p < .05$) compared with the integrated group (4.73 rating, $p < .05$), $d = 0.55$, and the split-attention group (7.17 rating, $p < .05$), $d = 1.99$. Tukey post hoc tests also revealed that the integrated group reported a significantly lower level of perceived cognitive load than the split-attention group, $d = 1.32$, indicating a large effect size.

Performance measures. Two separate one-way ANOVAs were conducted on recall and transfer test performance scores to explore differences between the three groups involved in the transfer task. Means and *SD*s for recall and transfer test scores are shown in Table 1.

A one-way ANOVA for recall scores revealed a significant effect of group; $F(2, 120) = 63.825$, $p < .05$, effect size $\eta_p^2 = 0.515$. Mean recall and transfer scores showed that the guided self-managed group had higher scores than the integrated group, which also had higher scores than the split-attention group. Post hoc comparisons using Tukey contrasts showed that the guided self-managed group performed significantly better than the other two groups. The integrated group performed significantly better than the split-attention group.

The one-way ANOVA for transfer questions also demonstrated a significant main effect of group; $F(2, 120) = 60.721$, $p < .05$, effect size $\eta_p^2 = 0.503$. Post hoc comparisons using Tukey contrasts showed that the guided self-managed group outperformed both the split-attention and integrated group. The integrated group also performed significantly better than the split-attention group.

Mental effort rating on the test. A one-way ANOVA was conducted on the instructional rating (of mental effort required) that the participants were asked to provide after the learning phase and after answering every question. The means and *SD*s for recall and transfer mental effort rating for the test phase are shown in Table 1. Results indicated significant effect of group on recall items; $F(2, 120) = 75.477$ $p < .05$, effect size $\eta_p^2 = 0.557$. Mean recall and transfer ratings showed that the guided self-managed group reported lowest levels of cognitive load, followed by the integrated group and the split-attention group.

Post hoc comparisons using Tukey contrasts showed that the guided self-managed group reported a significantly lower mental

effort (2.71 rating, $p < .05$) than the split-attention group (6.56 rating, $p < .05$), $d = 2.47$ and integrated group (4.63 rating, $p < .05$), $d = 1.71$. In turn the integrated group reported significantly lower mental effort than the split-attention group, $d = 0.80$, indicating a large effect size.

One-way ANOVA for transfer test items revealed a significant main effect of group $F(2, 120) = 85.925$, $p < .05$, effect size $\eta_p^2 = 0.589$. Post hoc comparisons using Tukey contrasts showed that the guided self-managed group reported lower mental effort (2.33 rating, $p < .05$) than the split-attention group (6.05 rating, $p < .05$), $d = 2.86$ and the integrated group (3.89 rating, $p < .05$), $d = 1.24$. The integrated group (3.89 rating, $p < .05$) reported significantly lower mental effort than the split-attention group (6.05 rating, $p < .05$), $d = 1.61$.

Guidance compliance. Results of the compliance measures for the transfer task indicated that 41 of the 42 participants (98%) followed the guidance about how to self-manage split attention. Participants were considered compliant if they performed at least one of the tasks provided that were highlighting material with a highlighter, using arrows to link text and diagram, underlining material, or marking circles on key words with a pencil or pen.

The most common strategy used by the participants (90%) was highlighting, underlining or numbering. The second most used strategy was drawing an arrow to link it to its corresponding place on the diagram. Only 43% of the participants drew a circle around the information. The use of at least these tasks is seemingly quite useful in understanding the instructional material. The level at which these tasks were conducted suggests that full utilization of the guidance contributed to higher performance scores.

Summary of Part 2 of the Experiment

The finding of higher transfer performance scores by students in the guided self-managed group compared with those in the conventional split-attention group was clearly evident. The superiority of the guided self-managed group might have resulted from the implementation of the guidance on how to integrate text and diagrams before learning the instructional material. The requirement to mentally integrate text with relevant aspects of the diagram by the split attention group, which had no guidance, might have contributed to poor performance by the split attention group. The results of performance scores on transfer items are similar to those found by Roodenrys et al. (2012) and Tindall-Ford et al. (2015). Such superior performance had been demonstrated, for example by Roodenrys et al. (2012), with Australian students studying educational psychology, by showing slightly increased accuracy of students in the guided self-management group over the integrated group with transfer test items.

For compliance measures, more than 94% followed the guidance offered. The performance of the guided self-management group may be attributed to the guidance given during the learning phase. The guidance to the self-management group improved performance across the two performance measures of recall, transfer and reported low cognitive load.

Part 2 of the Experiment was designed to follow up on the results observed in Part 1 of the Experiment and test whether participants would spontaneously transfer self-management skills to new and different split-attention instructional materials. If there is skills transfer, would this lead to a reduction in extraneous load and therefore enhance performance? Another important question is if skills transfer occurred, can this then be termed a self-management strategy? We contend that if learners are able to remember a skill and remember when to use it on their own, initial instruction was successful to automatize the self-management skills.

Students in the guided self-management group demonstrated higher performance than students in the split-attention for both recall and transfer tasks. These results again demonstrate the robustness of the split-attention effect. The participants in the guided self-managed group self-managed before attempting to learn the materials that improved their performance on test items. Participants in the integrated and split-attention groups who learned the same material but had no self-management knowledge performed worse that the guided self-managed group across all performance measures. The split-attention effect is further showed by the results of the integrated group that had higher performance scores than the split-attention group.

With regard to cognitive load in the transfer task, students in the guided self-managed instructional group, contrary to expectations, reported lower perceived cognitive load than students in the integrated group. In turn students in the integrated group reported lower perceived cognitive load than students in the split-attention group. Apparently, the processes required to work during the test phase demanded different amounts of mental effort in all conditions. When the data is differentiated for recall and transfer, the results still revealed the same tendency with the split attention group and integrated groups reporting higher levels of cognitive load.

The results of cognitive load do not support the hypothesis that a cognitive structure resulting from guided self-management instruction improves learning over one resulting from instruction emphasizing a conventional split-attention format. The high cognitive load experienced by the split-attention group resulted in the group investing less effort in more relevant learning processes, consequently performing poorly in recall and transfer tasks.

General Discussion

The aim of this study was to investigate the effect novice students' guided self-management by physically manipulating paper-based instructional materials on learning accounting. This was examined in Part 2 of the Experiment with different accounting materials.

The major finding from this study relates to the students' ability to learn to manage cognitive load created by instructional material that requires them to split their attention between diagram and text. As a precursor to demonstrating the self-management skills, it was necessary to demonstrate that the accounting instructional materials do indeed demonstrate such split-source format and that this has a negative effect on learning. Both studies presented in this study showed that when split attention was managed by the students by integrating text and diagrams, students consistently outperformed those in the split-attention and integrated groups.

In terms of test performances, the findings of the experiment strongly supported the hypothesis that students learn better when guided to self-manage instructional material rather than learning content under split-attention. Therefore, the evidence of split-attention within the learning materials was established. The inte-

grated format group performed significantly better than the traditional split-attention format group in terms of recall and transfer performance tests. Findings showed that the guided self-managed format outperformed the integrated format on recall and transfer test items. In practice students are often confronted with split-attention materials. If they are able to integrate them, they can learn more. The students who have learned to integrate can do well on these tasks, whereas students who have worked with integrated examples do not know what to do.

In terms of mental effort (perceived cognitive load) experienced during learning, there were significant differences across the three groups. Participants studying the guided self-managed format consistently reported lower cognitive load than the split-attention format. Our hypotheses was not confirmed as students studying under the guided self-managed format reported significantly lower cognitive load than students studying under the integrated format group. Finally, students in the integrated format group reported lower levels of cognitive load than students in the split-attention format group. A possible explanation of the lower cognitive load in the guided self-managed group may be that the guidance provided by the self-management prompts may have affected the overall perception of cognitive load. Future studies could investigate the contribution of the different instructional components to the students' perceived cognitive load. The recently developed more detailed rating scales of Leppink, Paas, Van der Vleuten, Van Gog, and Van Merrienboer (2013; see also Leppink, Paas, Van Gog, Van der Vleuten, & Van Merrienboer, 2014), could be used to disentangle these contributions.

The transfer of skills through self-management load techniques is a promising aspect that will enhance learners' performance in learning accounting. The present study reinforces the importance for instructors not just to design material according to CLT principles, but to present instructional formats in a way students can easily navigate.

A possible concern of the current research was the difference between the participants in the group taught to self-manage split attention and the other two groups. In Part 1 of the Experiment the students in the self-management group were guided on how to move text close to a diagram. In Part 2 of the Experiment the students in the guided self-management group were expected to use this guidance before answering the questions. It is possible that this difference in instruction contributed to the difference in performance between the groups.

It should be noted that we used an overall measure of cognitive load, which is of restricted informational value, because it does not allow the experimenter to conclude on the different types of cognitive load. Future studies could use the more detailed measures developed by Leppink et al., (2013; see also Leppink et al., 2014). Another limitation of this study is that a task may feel difficult for a variety of reasons other than cognitive load, such as difficulty in finding and combining the appropriate or selected strategies.

Implications for Instructors and Learners

In formulating instructional prescriptions from the present research, we took into consideration previous findings such as those by Roodenrys et al. (2012), who concluded that managing the split-attention format enhances performance. Roodenrys et al. (2012) also revealed the potential of students enhancing their performance by self-management of diagram and texts. Other studies have reported the deleterious effect of the split-attention effect on learning (e.g., Roodenrys et al., 2012; Tindall-Ford et al., 2015). It is important to note that in the current study, the benefit of integrated material was even more apparent when participants were instructed to integrate the materials for themselves.

In light of these recent findings, the results of the present study support the conclusion that instruction with emphasis on self-management of the instructional material is an appropriate alternative to conventional split-attention instruction. At the same time, caution is needed when considering this strategy since several studies have concluded that the self-management strategy has to be carefully implemented to enhance learning (e.g., Roodenrys et al., 2012; Tindall-Ford et al., 2015). For example, guidance on self-management has to be provided for instructional materials that are designed to evoke split-attention that cannot be understood in isolation. The guided self-management techniques learners may utilize involve numbering, linking text with diagrams using arrows, or highlighting keywords or concepts.

Many of the learning activities that novice undergraduate accounting students engage with in the classroom, whether related to reading, calculations, or other areas of studying accounting, impose considerable burdens on the limited capacity of working memory. These activities often require a student to hold in mind some information (e.g., a text) while attempting to match with relevant parts of a diagram. This is something that cognitive load theory and this study argue to be mentally challenging and not facilitating learning, yet these are the kinds of activities that novice learners struggle with. Therefore, instructors may need to design instructional materials that are already integrated or, maybe, guide students with crucial information to self-manage by properly integrating relevant study material to facilitate learning as presented in this study.

While this study enables us to suggest ways in which instructors can help learners achieve greater success early in their accounting undergraduate courses, it may also assist learners to solve problems by manipulating instructional materials by themselves. Students can be taught how to navigate through accounting lectures, studying for examinations and various other learning activities using guided self-management skills.

Instructional Implications for Textbook Writers

Numerous examples exist of instructional material not designed according to cognitive load theory principles in the area of introductory accounting. As illustrated in this study, students often encounter instructional material such as the statement of financial position (balance sheet), statement of cash flows, and statement of changes in equity. The general format involves a diagrammatic representation with text below or above the diagrammatic representation. An alternative instructional presentation would be to have the text embedded in the diagrammatic presentation that is referred to as integrated; and, as this study has shown, this would reduce the extraneous cognitive load and enhance learning.

Another conventional way of presentation, again using the example of a balance sheet, is to visualize the balance sheet in the form of an equation. The equation explained within the text would be that total assets equals liabilities plus owners' equity. Looking at the equation in this way shows how assets were financed; either by borrowing money (liability) or by using the owners' money (owners' or shareholders' equity). However, most balance sheets do not have the equation

depicted, and students are usually forced to have a mental representation of the equation in their minds as they try to make sense of the assets in one section and the liabilities and net worth in the other section that would make the sections "balance." Such type of presentation exerts an unnecessary load on working memory.

Conclusion

In conclusion, we think that the results of our work are promising for an underdeveloped area of "guided self-management effect." Like Roodenrys et al.'s work (Roodenry et al., 2012), we found that the split attention has a negative effect on learning. In addition we found that when learners self-manage instructional material it enhances learning. Theoretically, this has important implications with regard to evidence that learners can be instructed on how to self-manage rather than relying on the instructor to keep on reorganizing deficient instructional formats such as split attention formats into more effective formats such as integrated formats. From a practical perspective, these results are important for students, instructors and textbook writers. For students who frequently encounter split-attention learning material they can take control of their own cognition and learning. The results of this study reinforce the importance of instructors to present instructional formats in a way that students can easily navigate for guided self-management. Despite the potential revealed by these studies, for guided self-management to be successful, the onus still rests with the instructor to guide students to manage the load. At the same time, the research also raises new questions about the need for further research to establish the robustness of the student-led guided self-management effect and finding other methods students can be instructed to use to integrate separate text and diagrams to enhance learning. Future research is also needed to find other methods students can be instructed to use to integrate separate text and diagrams to enhance learning.

References

Agostinho, S., Tindall-Ford, S., & Bokosmaty, S. (2014). Adaptive diagrams: A research agenda to explore how learners can manipulate online diagrams to self-manage cognitive load. In W. Huang (Ed.), *Handbook of human centric visualization* (pp. 529–550). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-7485-2_21

Agostinho, S., Tindall-Ford, S., & Roodenrys, K. (2013). Adaptive diagrams: Handing control over to the learner to manage split-attention online. *Computers & Education, 64,* 52–62. http://dx.doi.org/10.1016/j.compedu.2013.01.007

Aleven, V., McLaren, B. M., Roll, I., & Koedinger, K. R. (2006). Toward meta-cognitive tutoring: A model of help-seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education, 16,* 101–130.

Ayres, P., & Paas, F. (2012). Cognitive load theory: New directions and challenges. *Applied Cognitive Psychology, 26,* 827–832. http://dx.doi.org/10.1002/acp.2882

Ayres, P., & Sweller, J. (2005). The split-attention principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 135–146). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511816819.009

Baddeley, A. (1992). Working memory. *Science, 255,* 556–559. http://dx.doi.org/10.1126/science.1736359

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 570–585. http://dx.doi.org/10.1037/0278-7393.33.3.570

Berthold, K., & Renkl, A. (2009). Instructional aids to support a conceptual understanding of multiple representations. *Journal of Educational Psychology, 101,* 70–87. http://dx.doi.org/10.1037/a0013247

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to cnhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, J. R. Pomerantz, M. A. Gernsbacher, R. W. Pew, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). New York, NY: Worth Publishers.

Bjork, R. A., & Kroll, J. F. (2015). Desirable difficulties in vocabulary learning. *The American Journal of Psychology, 128,* 241–252. http://dx.doi.org/10.5406/amerjpsyc.128.2.0241

Blayney, P., Kalyuga, S., & Sweller, J. (2010). Interactions between the isolated-interactive elements effect and levels of learner expertise: Experimental evidence from an accountancy class. *Instructional Science, 38,* 277–287. http://dx.doi.org/10.1007/s11251-009-9105-x

Bodemer, D., Ploetzner, R., Feuerlein, I., & Spada, H. (2004). The active integration of information during learning with dynamic and interactive visualisations. *Learning and Instruction, 14,* 325–341. http://dx.doi.org/10.1016/j.learninstruc.2004.06.006

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21,* 279–283. http://dx.doi.org/10.1177/0963721412452728

Choi, H. H., Van Merriënboer, J. J., & Paas, F. (2014). Effects of the physical environment on cognitive load and learning: Towards a new model of cognitive load. *Educational Psychology Review, 26,* 225–244. http://dx.doi.org/10.1007/s10648-014-9262-6

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review, 3,* 149–210. http://dx.doi.org/10.1007/BF01320076

Clark, R. C., & Mayer, R. E. (2008). *E-Learning and the science of instruction.* San Francisco, CA: Pfeiffer.

Clark, R., Ngyuen, F., & Sweller, J. (2006). *Efficiency in learning: Evidence based guidelines to manage cognitive load.* San Francisco, CA: Pfeiffer.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. http://dx.doi.org/10.3758/BF03193146

Florax, M., & Ploetzner, R. (2010). What contributes to the split-attention effect? The role of text segmentation, picture labelling, and spatial proximity. *Learning and Instruction, 20,* 216–224. http://dx.doi.org/10.1016/j.learninstruc.2009.02.021

Gall, M. D., Borg, W. R., & Gall, J. P. (2003). *Educational research: An introduction* (7th ed.). White Plains, NY: Longman Publisher.

Ginns, P. (2005). Meta-analysis of the modality effect. *Learning and Instruction, 15,* 313–331. http://dx.doi.org/10.1016/j.learninstruc.2005.07.001

Ginns, P. (2006). Integrating information: A meta-analysis of the spatial contiguity and temporal contiguity effects. *Learning and Instruction, 16,* 511–525. http://dx.doi.org/10.1016/j.learninstruc.2006.10.001

Goolkasian, P., Foos, P. W., & Eaton, M. (2009). Modality effects in sentence recall. *The Journal of General Psychology, 136,* 205–224. http://dx.doi.org/10.3200/GENP.136.2.205-224

Gordon, C., Tindall-Ford, S., Agostinho, S., & Paas, F. (2016). Learning from instructor-managed and self-managed split-attention materials. *Applied Cognitive Psychology, 30,* 1–9. http://dx.doi.org/10.1002/acp.3159

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38,* 23–31. http://dx.doi.org/10.1207/S15326985EP3801_4

Leahy, W., Chandler, P., & Sweller, J. (2003). When auditory presentations should and should not be a component of multimedia instruction. *Applied Cognitive Psychology, 17,* 401–418. http://dx.doi.org/10.1002/acp.877

Leppink, J., Paas, F., Van der Vleuten, C. P., Van Gog, T., & Van Merriënboer, J. J. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods, 45,* 1058–1072. http://dx.doi.org/10.3758/s13428-013-0334-1

Leppink, J., Paas, F., Van Gog, T., van Der Vleuten, C. P., & Van Merrienboer, J. J. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction, 30,* 32–42. http://dx.doi.org/10.1016/j.learninstruc.2013.12.001

Liu, T. C., Lin, Y. C., Tsai, M. J., & Paas, F. (2012). Split-attention and redundancy effects on mobile learning in physical environments. *Computers & Education, 58,* 172–180. http://dx.doi.org/10.1016/j.compedu.2011.08.007

Mason, L., Caterina, M. T., & Pluchino, P. (2013). Do fourth graders integrate text and picture in processing and learning from an illustrated science text? Evidence from eye-movement patterns. *Computers & Education, 60,* 95–109. http://dx.doi.org/10.1016/j.compedu.2012.07.011

Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist, 32,* 1–19. http://dx.doi.org/10.1207/s15326985ep3201_1

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511811678

Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 47–62). New York, NY: Macmillan.

Morrison, B. B., Dorn, B., & Guzdial, M. (2014). Measuring cognitive load in introductory CS: Adaptation of an instrument. In *Proceedings of the tenth annual conference on international computing education research* (pp. 131–138). New York, NY: ACM. http://dx.doi.org/10.1145/2632320.2632348

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skills in statistics: A cognitive load approach. *Journal of Educational Psychology, 84,* 429–434. http://dx.doi.org/10.1037/0022-0663.84.4.429

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38,* 1–4. http://dx.doi.org/10.1207/S15326985EP3801_1

Paas, F., & Van Gog, T. (2006). Optimising worked example instruction: Different ways to increase germane cognitive load. *Learning and Instruction, 16,* 87–91. http://dx.doi.org/10.1016/j.learninstruc.2006.02.004

Paas, F., Van Gog, T., & Sweller, J. (2010). Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational Psychology Review, 22,* 115–121. http://dx.doi.org/10.1007/s10648-010-9133-8

Paas, F., & Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86,* 122–133. http://dx.doi.org/10.1037/0022-0663.86.1.122

Paivio, A. (1986). *Mental representation: A dual coding approach.* Oxford, England: Oxford University Press.

Plass, J. L., Moreno, R., & Brünken, R. (2010). *Cognitive load theory.* New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511844744

Rau, M. A. (2015). Enhancing undergraduate chemistry learning by helping students make connections among multiple graphical representations. *Chemistry Education Research and Practice, 16,* 654–669. http://dx.doi.org/10.1039/C5RP00065C

Roodenrys, K., Agostinho, S., Roodenrys, S., & Chandler, P. (2012). Managing One's own cognitive load when evidence of split-attention is present. *Applied Cognitive Psychology, 26,* 878–886. http://dx.doi.org/10.1002/acp.2889

Rose, J. M., & Wolfe, C. J. (2000). The effects of system design alternatives on the acquisition of tax knowledge from a computerized tax decision aid. *Accounting, Organizations and Society, 25,* 285–306. http://dx.doi.org/10.1016/S0361-3682(99)00048-3

Samur, Y. (2012). Redundancy effect on retention of vocabulary words using multimedia presentation. *British Journal of Educational Technology, 43,* E166–E170. http://dx.doi.org/10.1111/j.1467-8535.2012.01320.x

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction, 22,* 129–184. http://dx.doi.org/10.1207/s1532690xci2202_1

Seufert, T. (2003). Supporting coherence formation in learning from multiple representations. *Learning and Instruction, 13,* 227–237. http://dx.doi.org/10.1016/S0959-4752(02)00022-1

Seufert, T., & Brünken, R. (2006). Cognitive load and the format of instructional aids for coherence formation. *Applied Cognitive Psychology, 20,* 321–331. http://dx.doi.org/10.1002/acp.1248

Stern, E., Aprea, C., & Ebner, H. G. (2003). Improving cross-content transfer in text processing by means of active graphical representation. *Learning and Instruction, 13,* 191–203. http://dx.doi.org/10.1016/S0959-4752(02)00020-8

Sweller, J. (2006). The worked example effect and human cognition. *Learning and Instruction, 16,* 165–169. http://dx.doi.org/10.1016/j.learninstruc.2006.02.005

Sweller, J. (2015). In Academe, what is learned, and how is it learned? *Current Directions in Psychological Science, 24,* 190–194. http://dx.doi.org/10.1177/0963721415569570

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory.* New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4419-8126-4

Tindall-Ford, S., Agostinho, S., Bokosmaty, S., Paas, F., & Chandler, P. (2015). Computer-based learning of mathematical concepts by studying instructor-managed or self- managed split-attention materials. *Journal of Educational Technology & Society, 18,* 89–99.

Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist, 43,* 16–26. http://dx.doi.org/10.1080/00461520701756248

van Merriënboer, J. J. G., & Ayres, P. (2005). Research on cognitive load theory and its design implications for e-learning. *Educational Technology Research and Development, 53,* 5–13. http://dx.doi.org/10.1007/BF02504793

van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17,* 147–177. http://dx.doi.org/10.1007/s10648-005-3951-0

Weygandt, J. J., Chalmers, K., Mitrione, L., Fyfe, M., Kieso, D. E., & Kimmel, P. D. (2010). *Principles of financial accounting.* Australia: Wiley Australia Ltd.

Yue, C. L., Bjork, E. L., & Bjork, R. A. (2013). Reducing verbal redundancy in multimedia learning: An undesired desirable difficulty? *Journal of Educational Psychology, 105,* 266–277. http://dx.doi.org/10.1037/a0031971

Received August 10, 2015
Revision received March 15, 2016
Accepted March 15, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 2, 233–244

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000131

Learning to Read With and Without Feedback, In and Out of Context

Sandra Martin-Chang
Concordia University

The self-teaching hypothesis posits that enduring orthographic and phonological representations are produced when children independently recode print into speech. However, very little research has examined how children self-teach when initial decoding attempts are weak or ineffective. In this within-participant design, 25 students in Grade 2 learned to read 85 different words in 4 conditions. Words were read in and out of context, with and without feedback. Accuracy rates were recorded throughout 5 training sessions (2 word repetitions per session = 10 repetitions in total). A posttest was administered after a 6-day delay by reinstating the training materials. At the end of training, the highest accuracy scores were observed when children read in context/feedback followed by when they read in isolation/feedback, and then in context/no feedback; the lowest accuracy scores were observed when children read in isolation/no feedback. This pattern remained over the retention period, suggesting that external support from feedback, and top-down support from context, can help create word representations in memory. The results are discussed in relation to the importance of whole-word phonology within self-teaching.

Keywords: context, decoding, feedback, isolation, self-teaching, word reading

Share (2004) noted that children "self-teach" the majority of the words they can read. The ability to self-teach is associated with two potential factors. The first factor relates to the *process* of recoding; namely, focusing on grapheme-to-phoneme correspondences during decoding may help to create well-specified orthographic representations in memory (Share, 1995). The second factor relates to the *product* of recoding; in this case, focusing simultaneously on whole-word orthography and phonology, may help amalgamate written words with their spoken pronunciations (Elbro, de Jong, Houter, & Nielsen, 2012). When decoding is successful, it is difficult to disentangle the effects of these two factors: proficient decoding results in the word's correct pronunciation. However, ineffective decoding creates the opportunity to examine the second possible factor more closely. For example, when decoding skills are weak, or the word to be read is difficult, correct pronunciations are more likely to be activated when children read in context. Moreover, when decoding fails to produce the spoken word altogether, correct pronunciations can be provided via feedback from a "teacher." The current experiment explored the second potential factor involved in self-teaching by examining children's growth in reading accuracy as they read in, and out, of context—with, and without, corrective feedback.

This article was published Online First June 2, 2016.
I gratefully acknowledge support from the Social Sciences and Humanities Research Council (SO1294) and the children, teachers and principals who participated in this study. I also thank Gene Ouellette, Kyle Levesque, and Holly Recchia for their thoughts and comments on an earlier draft of this article.
Correspondence concerning this article should be addressed to Sandra Martin-Chang, Department of Education, Concordia University, 1455 de Maisonneuve Boulevard West, Montreal, Quebec, Canada, H3G 1M8. E-mail: smartinc@education.concordia.ca

The Self-Teaching Hypothesis

The notion behind self-teaching is that children build up "sight word" lexicons—words that can be read automatically—by creating orthographic representations as they decode words. In essence, "sounding out" new words induces a form of cognitive processing akin to "focal attention" (Samuels, 1967) because children's attention is focused on the letters, letter patterns, and letter sequencing that make up each word's unique orthographic form (Share, 1999). This process allows children to gradually accumulate general orthographic knowledge (knowledge about the language as a whole), by internalizing orthographic representations of many specific words (e.g., slowly coming to understanding that the "ss" spelling pattern is legal at the end, but not beginning of English words through exposure with words such as *less* and *sell*, *boss*, *sob*, etc.). Support for this hypothesis has come from several empirical investigations showing robust word-specific orthographic gains after a minimal number of successful reading experiences for both nonwords (Bowey & Muller, 2005; Cunningham, Perry, Stanovich, & Share, 2002; Ouellette, 2010; Ouellette & Fraser, 2009; Share, 1999; Wang, Castles, Nickels, & Nation, 2011) and, to a lesser extent, real words (Cunningham, 2006; Landi, Perfetti, Bolger, Dunlap, & Foorman, 2006).

Ziegler, Perry, and Zorzi (2014) recently simulated the self-teaching process by building a connectionist model around two main assumptions: (a) that young children begin to read with a well-developed spoken vocabulary; and (b) that before reading begins in earnest, children are explicitly taught phonics. The Phonological Decoding Self-Teaching Model quickly learned how to read more than 25,000 words and was able to generalize this learning to read nonwords. Critical to the self-teaching hypothesis, it was able to accomplish these tasks without the use of feedback from a teacher. Interestingly context was also seen to play a key role in the model. For example, when the correct word was generated as one of several possible word candidates, it was

chosen based on the premise that in "real learning situations with real texts, children have additional information from the story context, semantics, or syntax to help them choose the correct target" (p. 6). In short, given alternative pronunciations, Ziegler et al., suggest that the ability to select the correct pronunciation is attributed to additional information gained from reading in context.

Self-Teaching in Context

The self-teaching hypothesis is aligned with decades of research showing the importance of explicit and systematic teaching of letter-sound correspondences (Snowling & Hulme, 2011). Indeed, a core tenet of self-teaching is that children need to understand the alphabetic principle so that decoding can take place. The self-teaching hypothesis also acknowledges that most word knowledge is learned implicitly via experience with texts (e.g., Elgort, Perfetti, Rickles, & Stafura, 2015; Nagy, Anderson, & Herman, 1987).

Words in naturalistic text are difficult to predict (Gough, Alford, & Holley-Wilcox, 1981); therefore, when children read authentic texts they seldom rely on guessing from context as their primary strategy for reading (Nation & Snowling, 1998; Tunmer & Chapman, 1995). Still, under certain conditions predicting words from context becomes more likely. For example, when reading highly constrained texts (e.g., "Roses are red, violets are _"), individuals are less likely to fixate on the predictable words. Moreover, when the predicable words *are* fixated, the amount of time spent looking at them tends to be shorter in duration compared with those that are less predictable (Ehrlich & Rayner, 1981). Under normal circumstances, however, rather than simply "guessing" from context, children use both partial decoding attempts *and* the structure of the text to arrive at correct word pronunciations (Nation & Snowling, 1998; Share, 2004; Tunmer & Chapman, 2012). The advantage gained from this type of crosschecking between semantics and print has been termed "contextual facilitation."

The beneficial effects of reading in context have been well documented (Martin-Chang & Levy, 2006; Martin-Chang, Levy, & O'Neil, 2007; Roth & Perfetti, 1980; Stanovich, Nathan, West, & Vala-Rossi, 1985). What remains contentious is whether the performance gains observed during online contextual reading, accumulate to result in crystalized learning that is capable of supporting later reading fluency (for the distinction between online performance and crystallized learning see Byrne et al., 2013).

Landi and colleagues (2006) were interested in this distinction between reading performance and generalized learning. They identified two written word sets that children in Grades 1 and 2 (Experiment 1) were unable to accurately read aloud. Half of the words were then shown in isolation, and the other half were presented in predicable sentences. Landi et al. reported that children were able to read more words in predictable sentences than in isolation (13.29 words in context compared with 4.88 words in isolation). However, both sets of words were read with similar accuracy when presented in isolation one week later (6.15 words in context, 5.3 words in isolation). This led Landi et al. to conclude that children *performed* better in context initially, but that reading words in isolation was superior for *learning*. However, as described previously, very highly constrained text, which is generally not representative of children's more naturalistic contextual reading, encourages top-down processing and reduces the need for students to focus on the print. Therefore, Landi et al.'s selection of text may have inadvertently reduced the opportunities for self-teaching in context.

Cunningham's work (Cunningham, 2006) addressed some of these issues by inviting children in Grade 1 to read four coherent and four scrambled passages that were longer and less predictable than the materials used by Landi et al. (2006). Cunningham reported that words were read more accurately when they were initially presented in a meaningful context (83.6%) compared with in a scrambled passage (67%). However, 3 days after self-teaching, the children performed similarly on an orthographic choice task and a spelling task, regardless of whether the words were first read in context or isolation. A posttest measure of reading accuracy was not included—perhaps because there were only eight words to be learned in total (one word per passage). Therefore, the results of this study cannot be directly compared with those of Landi et al., in terms of word reading accuracy.

Self-teaching in and out of context was also explored in two experiments by Wang et al. (2011). The authors created a set of words that were understood verbally by associating nonwords with meanings. After the pronunciations and definitions of the nonwords were well understood orally, they were presented in writing. The spellings of the nonwords remained the same over two experiments; however, the pronunciations associated with the nonwords differed. In Experiment 1 the pronunciations were regular while in Experiment 2, they were irregular. Wang et al. reported that the regular words were read more accurately in context during the first self-teaching trial and that the irregular words were read more accurately across all four contextual training trials. They concluded that when decoding is difficult, as it is for irregular words, context helps children read words accurately. However, like Cunningham (2006), Wang et al. used a small learning set (four words per condition). Therefore, it is unclear whether their findings would generalize to reading longer passages.

Reviewing the work of Martin-Chang and Levy (2005) can help resolve some of these questions. They presented average readers in Grade 2 (Experiment 2) with 85 real words read in a meaningful story and 85 different words read in isolation. The authors found that the children read the target words more accurately in context compared with in isolation during training. They also found that children read *new stories* faster and more accurately if the words in the stories had first been trained in a different context. However, Martin-Chang and Levy departed from the methodology of other researchers by electing to give corrective feedback in response to children's errors; therefore, their results cannot be directly compared with the self-teaching literature (e.g., Cunningham, 2006; Landi et al., 2006; Wang et al., 2011).

Self-Teaching and Feedback

A defining characteristic of the self-teaching model is that feedback from external sources is not required for word acquisition. However, very little research has compared how self-teaching in the absence of an expert compares to learning with the assistance of feedback.

When the child receives whole word feedback, the "whole word" is supplied after an error. This type of feedback is often negatively contrasted with graphophonemic feedback, which relates individual letters or letter clusters with specific sounds. Whole word feedback—also termed "terminal feedback"—has

been faulted for ending the children's decoding attempts, (Evans, Barraball, & Eberle, 1998). Phonological recoding has been speculated to be a crucial step in learning how to read words fluently (Share, 2004); consequently, if receiving assistance *discourages* children from actively recoding, or ends the recoding process prematurely, it might also impair the quality of word representations in memory. Indeed, Landi and colleagues (Landi, 2013; Landi et al., 2006) have speculated that students will be less likely to spend cognitive resources during initial decoding attempts when they know the correct pronunciation is forthcoming. If this is the case, feedback would be expected to reduce exhaustive grapheme-phoneme recoding, to the detriment of self-teaching.

Without question, a teacher or parent who "supplies" a misread word is doing the important work of recoding the symbols into speech on the child's behalf; yet it is possible that hearing the pronunciation while attending to the print might put the child in a better position to phonologically recode the word on subsequent trials. As argued by Ehri (2014), when "readers see a new word and say or hear its pronunciation, its spelling becomes mapped onto its pronunciation and meaning" (Ehri, 2014, p. 6). Barbetta, Heward, Bradley, and Miller (1994) provided evidence to support this view by teaching five students in Grade 2 to read words in isolation with immediate or delayed whole-word feedback. The results showed that immediate feedback was more profitable than delayed feedback in the acquisition and maintenance of word reading. The authors speculated that the immediate feedback reduced the likelihood that the same mistakes were repeated throughout training. Additionally, immediate feedback allowed the pronunciation to be heard soon after the print was seen, which may have contributed to the amalgamation of the words' phonological and orthographic forms.

Current Investigation

The literature discussing self-teaching stresses the importance of each "successful recoding experience" as if it was one unified entity (Cunningham, 2006; Share, 1999, 2004). However, within every fruitful decoding attempt there are two factors that could be contributing to self-teaching: (a) grapheme-phoneme recoding; and (b) the pairing of whole word orthography and phonology. This pairing of complete spoken words with their respective letter strings could be achieved in at least three ways: by pure bottom-up decoding (the first factor in self-teaching); by decoding that is supplemented with top-down support from context; or by feedback that is provided by an external source after a failed decoding attempt. The question is, does providing support (from context or feedback) weaken or strengthen long-term word recognition? If independent grapheme-phoneme recoding is critical to self-teaching, then situations where there is very little support (e.g., reading in isolation without feedback) would be expected to produce the highest degree of accuracy over training. In contrast, if pairing whole-word orthography and phonology is central to creating word representations in memory (Ehri, 2014), then situations that offer the most support for reading accuracy (e.g., reading in context with feedback) should result in superior accuracy. The current investigation tested these hypotheses by having students read a large set of words, in and out of meaningful text, with and without feedback.

Method

Design

A within-subjects design was implemented with two experimental manipulations: the availability of context and the provision of feedback. Specifically, the first manipulation involved whether the target words were presented in *isolation* or in *context*. The second manipulation involved whether whole-word *feedback* was provided or whether *no feedback* was given. Taken together, the training phase of the study consisted of four distinct experimental reading conditions: *context/feedback*, *isolation/feedback*, *context/no feedback*, and *isolation/no feedback*. Participants were exposed to four unique sets of target words, which were counterbalanced across all four experimental conditions (see Figure 1). The experiment was conducted in two 18-day blocks. In one block, the children were given feedback during both the context and isolated word training conditions. During the other block, the children did not receive feedback during either context or isolated word training. Each block contained five training sessions (on Days 1, 3, 7, 9, and 11). A posttest was administered on the last day of each block (Day 18) to determine if accuracy gains would be maintained over a delay.

Participants

Twenty-eight participants were recruited from three Grade 2 classrooms in central Canada. The children's teachers all reported

Block 1: Feedback*

Monday	Tuesday	Wednesday	Thursday	Friday
Screening	Day 1 Session 1: List A** Story B	Day 2	Day 3 Session 2: List A Story B	Day 4
Day 7 Session 3: List A Story B	Day 8	Day 9 Session 4: List A Story B	Day 10	Day 11 Session 5: List A Story B
Day 14	Day 15	Day 16	Day 17	Day 18 Retention: List A Story B

Block 2: No feedback*

Monday	Tuesday	Wednesday	Thursday	Friday
Screening	Day 1 Session 1: Story C** List D	Day 2	Day 3 Session 2: Story C List D	Day 4
Day 7 Session 3: Story C List D	Day 8	Day 9 Session 4: Story C List D	Day 10	Day 11 Session 5: Story C List D
Day 14	Day 15	Day 16	Day 17	Day 18: Retention Story C List D

Figure 1. Experimental training design, Story = context, and list = isolation. * The order of the feedback conditions and the training conditions was counterbalanced over all participants, so that half of the children received the no-feedback training condition first, and the other half received the story condition first. In addition, the training words were counterbalanced over all conditions so that each set of words was trained in each of the four conditions. ** Each list and story contained two repetitions of 85 different training words. Therefore, the children read a total of 170 unique words (half in a story, half in a list), twice, each day.

using strong phonics training (e.g., Jolly Phonics) in their regular classroom instruction. Children with parental consent were screened with the reading subtest of the Wide Range Achievement Test—Third Edition (WRAT3; Wilkinson, 1993). In total, three participants were omitted from the study. The first student had a low standardized WRAT score (<70), the second was absent for an extended period of time, and the third was shy and preferred not to read aloud. The final sample consisted of 25 students (17 girls and 8 boys). Participants were in Grade 2 and were, on average, 7 and half years old (M = 7 years and 7 months, range = 6 years and 7 months to 8 years and 3 months). The average standardized WRAT score was 97 (SD = 14, range from 75 to 130). After each session, the children were thanked with small gifts (e.g., stickers, pencils). They were also invited to choose an age appropriate book to take home after each 18-day block.

Materials

To use longer and more complex passages for the children, four nonoverlapping sets of 85 real words were counterbalanced over the four training conditions (340 unique words in total/4 conditions = 85 words per condition). The words were selected because they were contained in 7-to-8-year-olds' spoken vocabularies (Martin-Chang & Levy, 2005). For the present study, six Grade 2 teachers (including a classroom teacher from the current sample) also vetted the materials to confirm that they contained words children could understand orally. The teachers indicated that on average, only 3.6% of the 340 words might be difficult for children in Grade 2 to understand; the remaining 96.4% of the words were deemed to be common in children's spoken vocabularies.

The materials included words with both regular and irregular spellings and words with a range of morphological complexities (see Appendix A). The lists were not equated in terms of spelling patterns or orthographic complexity; however, the average number of letters (average = 5.7 letters, range 5.5 letters to 5.9 letters) and morphemes (average = 1.44 morphemes, range 1.4–1.5) per list were controlled for. The items were not screened to avoid additional word exposures (see also Cunningham, 2006); therefore, it was difficult to track how many of the words could already be read before the first session.

The isolation reading condition was achieved by presenting the words individually on a computer screen. The target items were repeated twice within each isolated-word list to produce a total of 170 target words. To create the context condition, four narratives were created around each of the lists (see Appendix B for an example). The target words were repeated twice within each training story to produce a total of 170 target words. The four stories contained some written words (approximately 30–40%) that were above the children's reading ability and were all evaluated to be a 4.0 grade level by the Flesh- Kincaid readability test. They were written to resemble children's passages and ranged from 675 to 766 words in length. As would be the case within authentic texts, there was differing amount of contextual support for different words.

Procedure

Two training conditions were conducted during each 18-day block (see Figure 1). It was speculated that receiving feedback on only some trials might create unnecessary confusion for the children (e.g., they might wait for feedback on trials where no feedback was forthcoming). Therefore, the feedback manipulation was administered in separate blocks. The order of the blocks was counterbalanced so that half of the children received feedback in both conditions (context and isolation) during the first block and the other half of the children received feedback in both conditions (context and isolation) during the last block. Within each block, the order of the conditions was counterbalanced so that context was presented first for half of the children and presented last for the rest. It is important to note that, while the context and isolation manipulation occurred on the same days (either in the feedback or no feedback blocks), each of the four reading conditions corresponded to different word sets. The delayed posttest occurred on Day 18 of each block.

Two trained research assistants worked with the children in a quiet place in their school. The first experimenter conducted the training phase, while a second experimenter, who was blind to the feedback condition used during training, conducted the posttest.

Training phase. The students were given praise and encouragement throughout the entire experiment (not contingent on correct reading). However, during the feedback block, the children were also given whole-word corrective feedback after reading errors. In the no feedback block, the participants were asked to read as if they were alone; here, they were not given assistance of any kind. Within each block (feedback/no feedback), the children read two material sets; each set included 85 different words. The children read one story (context) and one list (isolation) during each training session. The stories and lists contained two repetitions of each word.

During isolated-word training, individual words were presented on a computer screen for 2 s, followed by a fixation point. The rapid pace of word presentation kept the task from becoming overly long and/or monotonous and the frequent encouragement and/or feedback from the experimenter kept the task from feeling completely solitary. To ensure that the lists did not appear too "story like" (i.e., not rapid serial visual presentation) the words were shown in a fixed- randomized order, with the only stipulation being that no word was presented twice in a row (85 words $\times$ 2 repetitions = 170 words). If the children were in the feedback block, the experimenter provided whole-word feedback after errors. The children were not asked to repeat the corrected word. Pauses longer than 2 s were considered omissions. Inaccurate attempts and omissions were both marked as errors. If the child was in the no feedback block, training continued without interruption. The experimenter discreetly coded the answer as correct or incorrect during the fixation cross, and prompted the onset of the next word by pressing a computer key. The sessions were audio recorded so that the scoring sheets could be verified for accuracy.

During context training, a shared reading paradigm was adopted where the participants read only the target words (85 $\times$ 2 repetitions = 170 words per session). The shared reading paradigm equated the task demands of the training conditions (isolation and context). The children were asked to follow along with the story and read the words that were bolded and underlined while the experimenter read the remainder of the story. This style is similar to one that might be adopted by a parent who pauses to let the children read some of the words in a "daily reading" story at home. The experimenter read at a natural pace, pausing at each target

word. If the child was in the feedback condition, the experimenter would read the target word after the child made an error or failed to attempt the word after 2 s. The story continued without the child repeating the corrected word. In the no-feedback condition, the experimenter would resume with the story after errors. Once again, the sessions were audiotaped for scoring purposes. The experimenter discreetly coded the child's reading as correct or incorrect as they were reading and inaccurate attempts and failures to respond were both marked as errors.

Testing phase. The delayed posttest occurred on the last day of the 18-day testing block. During the posttest, the children were asked to read the same materials that they had used during their training phase (i.e., same contextual story and same isolation word list) to determine if the reading accuracy gains from the training sessions had been retained. The only difference between the training phase and the posttest was that the children were not offered assistance, regardless of whether they had been in the feedback condition during training. If the child paused for more than 2 s, the experimenter reassured the child that it was fine to "skip it" or "keep going."

Results

Word Reading in Session 1

As shown in Figure 2, the children were able to read many of the target words at the onset of training. Accuracy ranged from 61 to 71% during the first session. To determine if there were any initial differences between the conditions, a 2 (context, isolation) × 2 (feedback, no feedback) repeated measures Analysis of variance (ANOVA) was conducted on the accuracy scores from the first training session. The significant main effect of context confirmed that children were able to read more words correctly in context compared with in isolation at the beginning of training, $F(1, 24) = 39.73$, $MSE = 4{,}121.64$, $p < .001$. The difference between word reading in context versus in isolation corresponded to a very large effect ($r = .79$). There was no main effect of feedback, indicating the accuracy scores in Session 1 were similar regardless of whether the children were in the feedback or no-feedback condition ($p = .15$). The Context × Feedback interaction was not significant ($p = .65$). Similar accuracy scores were expected in the two feedback

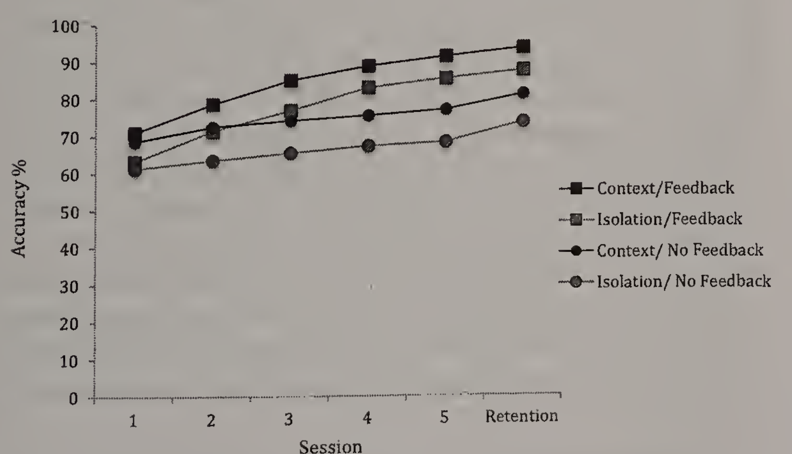


Figure 2. Percentage of words read correctly during the five training sessions and posttest.

conditions because the supplementary instruction had not had time to exert influence on training.

Word Reading in Sessions 1–5

The number of words that could be read inclusively from Sessions 1–5 was also analyzed. Figure 2 depicts the percentage of words read correctly during each session. The highest accuracy scores from Session 2 onward were noted when children read in context/feedback. The lowest scores were observed when they read in isolation/no feedback. Figure 2 also shows that when children read in isolation/feedback, they started training with reduced accuracy, but by Session 3, the children's scores in the isolation/feedback condition had surpassed the context/no-feedback condition.

To evaluate whether participants' word reading accuracy improved across training sessions, and, if so, whether this change was modulated by the experimental reading conditions, a 2 (context, isolation) × 2 (feedback, no feedback) × 5 (Sessions: 1–5) repeated measures ANOVA was conducted. Inspection of Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of Session, $\chi^2(9) = 103.25$, $p < .001$, and the interaction between Feedback and Session, $\chi^2(9) = 65.31$, $p < .001$. The degrees of freedom associated with these effects were corrected using Greenhouse-Geisser conservative estimates of sphericity ($\varepsilon = .31$ for the main effect of Session and $\varepsilon = .40$ for the Feedback × Session interaction; Field, 2009). Results from the ANOVA analysis confirmed significant main effects of context ($F(1, 24) = 35.51$, $MSE = 21{,}496.58$, $p < .001$, $r = .77$) and feedback ($F(1, 24) = 39.43$, $MSE = 37{,}243.48$, $p < .001$, $r = .79$), both of which corresponded to large effect sizes (Cohen, 1988). However, the Context × Feedback interaction was not significant ($p = .236$). Furthermore, the significant main effect of session showed that children were able to read a greater number of words as they participated in additional practice sessions, $F(1.26, 30.33) = 65.24$, $MSE = 31{,}879.71$, $p < .001$. Planned pairwise comparisons, with Bonferroni adjustments for multiple comparisons, confirmed that each session was associated with significantly greater learning in word reading compared with the performance in the previous session (all pairwise $ps < .003$). More importantly, this main effect was qualified by a significant Feedback × Session interaction, $F(1.58, 37.89) = 36.34$, $MSE = 5813.50$, $p < .001$, indicating that feedback resulted in accelerated learning over sessions.

Planned comparisons using the "repeated" contrast function was used to examine the interaction between Feedback and Session more carefully. Repeated contrasts are especially useful in a repeated-measures design in which the level of a variable has a meaningful order (e.g., Session 1, 2, 3, etc.; Field, 2009). In terms of the current study, the repeated contrasts compared the performance of each session to the performance at the previous session (Session 3 vs. 2), and evaluated the degree of learning in relation to the effects of feedback. At nearly every successive practice session, the Feedback condition was associated with significantly greater learning in word reading accuracy compared with the No Feedback condition (all $ps < .004$; Session 1 vs. 2, $r = .76$; Session 2 vs. 3, $r = .65$; Session 3 vs. 4, $r = .56$). Although there was a significant improvement in reading accuracy between Sessions 4 and 5 overall, the improvement was statistically similar for

the Feedback and No Feedback condition (Session 4 vs. 5, $p = .06$. The Context × Session ($p = .50$), and the Context × Feedback × Session interactions ($p = .19$) were not significant, indicating that the initial benefit of reading in context was maintained over all the training sessions.[1]

The contextual advantages observed throughout training raised the question of whether the effects were being driven by a small number of highly irregular items within the larger word set. To gain insight into this question, the data were reorganized to be examined from a within-word perspective. Simply put, this preliminary analysis shed light on how many times the *same word* was read more accurately (by different children) when it was presented in context compared with when it was presented in isolation. Reorganizing the data in this manner resulted in a less-than-ideal data set that was very large (85 words × 4 lists × 2 forms of feedback = 680 items) with very few observed values per item (approximately six participants per condition). Nevertheless, this exploratory inspection of the data revealed that when read with feedback, 215/340 words were read more accurately in context, 106/340 words were read more accurately in isolation, and 18/340 words were read equally accurately in context and isolation. A Wilcoxon signed-ranks test found these differences to be significantly different ($z = 7.324$, $p < .001$). An uncannily similar pattern was observed when the words were read without feedback: 213/340 words were read more accurately in context, 106/240 words were read more accurately in isolation, and 20/340 words were tied. Once again, a Wilcoxon signed-ranks test found these differences to be significantly different ($z = 8.202$, $p < .001$). Although the original study was not designed with these analyses in mind, the descriptive results suggest that the context effects are not being driven by a small number of words.

Delayed Posttest

On the 18th day of each training block, the children were once again presented with the training materials to measure retention of learning. This follow-up task was administered to determine if the reading gains made during training would be maintained over a 6-day period. As depicted in Figure 2, the general patterns favoring context and feedback were also observed during the posttest (context/feedback = 93.98%, isolation/feedback = 87.93%, context/no feedback = 81.45, isolation/no feedback = 73.89). Furthermore, the accuracy scores improved over the retention period, indicating that the children treated the delayed posttest like an additional training session. A 2 (context, isolation) × 2 (feedback, no feedback) × 2 (Session 5, posttest) repeated measures ANOVA confirmed that all three main effects were significant and corresponded to large effect sizes (Cohen, 1988): context ($F(1, 24) = 24.38$, $MSE = 7{,}267.92$, $p < .001$, $r = .71$), feedback ($F(1, 24) = 48.51$, $MSE = 30{,}582.28$, $p < .001$, $r = .82$), and time of test ($F(1, 24) = 7.69$, $MSE = 1{,}793.13$, $p = .011$, $r = .49$). The absence of any significant interactions (all $ps > .29$) suggested that the effects of context and feedback were similar at both the end of training and during the retention task.

Discussion

Successful decoding experiences contain two closely linked components, each of which could be contributing to the formation of orthographic representations in memory. The first factor involves the *act* of decoding, which can only function if the reader is attending to the orthographic details of the word. The second factor involves the *result* of decoding, which happens when the reader accesses the word's spoken pronunciation. Thus far, the lion's share of research on self-teaching has been dedicated to exploring the decoding process (the first factor), with fewer resources directed at examining the importance of matching whole-word phonology and orthography (the second factor). Granted, when decoding activates the correct spoken word, these two factors become tightly woven and any additional contribution beyond decoding is difficult to measure. However, when decoding is ineffective—as it was for 30 to 40% of the words read in Session 1—it creates an opportunity to experimentally bolster the second factor and measure resulting changes in word learning. In the current experiment, self-teaching was tracked while children read under two conditions—namely, context and feedback—hypothesized to increase the availability of word pronunciations. If the availability of whole-word phonology contributes to the creation of word specific representations in memory, then higher learning rates would be expected in these more supportive conditions.

When contemplating the two feedback conditions it becomes apparent that reading was more accurate when words were presented in context compared with in isolation. From the second session onward, the highest accuracy scores were observed when children read with the benefit of *both* context *and* feedback; this pattern remained unchanged over a 6-day delay. This may suggest that generating and/or saying the pronunciation aids children more than just hearing the correct pronunciation of the word after an error. Such an interpretation fits nicely with the work of Ehri, who has found accelerated learning rates when children were asked to read words out loud compared with silently (see Ehri, 2014 for review). Alternatively, it could suggest that children are simply more successful when reading in context, and that this advantage is maintained even with the additional support of feedback. The present experiment is not able to tease apart these alternative hypotheses, however, in either case, the data clearly show that training is more effective when children have access to the words' pronunciations, and that effortful grapheme-by-phoneme decoding is not the only process for creating word representations in memory.

Context also helped children read more accurately when they read without feedback. During the first no-feedback session, children read 12.36 more words in context than they did in isolation. This contextual benefit was largely maintained throughout the duration of training. Indeed, it took children five sessions reading in isolation without feedback (116.43 words) to achieve the same degree of accuracy as they had on the first day when reading in context without feedback (116.64 words). The benefits observed from context also remained stable over a 6-day delay. Framing these findings in terms of the two factors hypothesized to impact self-teaching helps to explain how the top-down support from context aids in long-term word learning. The contextual facilitation effect describes how children use partial word decodings and

[1] The same pattern of results was found when only words that were read incorrectly in Session 1 were included in the analysis in a 2 (context, isolation) × (feedback, no feedback) × 4 (Sessions 2–5) ANOVA.

contextual constraints to improve children's immediate word reading performance. The byproduct of increased reading accuracy is greater access to whole-word phonology and thereby, more opportunities to amalgamate speech and print, and hence, greater learning (Ehri, 2014; Share, 2004).

Thus far, it seems that reading with both forms of support (context and feedback) offers greater benefits than reading without either of these scaffolds (isolation without feedback). However, what if only one form of support was available? Do children gain more from reading in context (without feedback) or by reading with feedback (out of context)? This is an interesting question from a theoretical standpoint because of the fact that feedback pairs a word's verbal pronunciation with its written form on *every* repetition (either the child reads the word correctly, or the pronunciation is provided). In contrast, context allows a greater proportion of words to be read correctly initially, but children will not generate/hear the correct pronunciations of all of the words on every trial. Therefore, if seeing and hearing words in close succession is a driving force behind gains in accuracy across sessions, then receiving feedback should produce greater benefits than reading in context. This prediction was born out in the current study. Word reading in the isolation/feedback condition improved at a faster rate, and ultimately surpassed the context/no feedback reading scores.

The present study elected to use whole word feedback, which is generally regarded as less helpful than graphophonemic feedback during parent–child reading activities (Evans et al., 1998). The fact that marked reading improvements were found after even the *least* effective form of guidance was given, makes a convincing case for the use of feedback. It should be noted, however, that the children in the current study were in Grade 2 and they already had a basic foundation of decoding skills. Therefore, these results might not be replicated with a younger group of children.

Conjectures have been raised that *independently* generating phonology from print may provide ideal circumstances for long-term learning. A strong interpretation of this statement suggests that corrective feedback may, in fact, be disadvantageous to reading development (Landi, 2013). The data reported here do not support this hypothesis. Children experienced the most difficulty when they were reading in isolation without feedback. However, it is worth highlighting that even under the least supportive conditions, children were able to learn more than 12 words over and above the 104 words that were read correctly during the first session of isolated-word reading without feedback. Therefore, the current study adds to the literature (cf., Share, 1995, 1999, 2004) by showing that children are capable of independently using recoding skills to self-teach with the help of nothing more than pure decoding. However, they learned more than triple as many words at the end of training when they were given whole-word feedback in isolation compared with when they were left to read in isolation alone. These data suggest that hearing the spoken word soon after an unsuccessful reading attempt may help children form partial orthographic representations in memory that can be referenced and refined on future word encounters.

When contemplating the role of feedback in this investigation it is important to note that children were actively attempting to decode the words before hearing the word pronunciations. The same pattern of results would not be expected if the children did not yet understand the alphabetic principle or if they were simply "following along" while the teacher read the text. In this sense, even with feedback, the main components of self-teaching remain intact. That is, for feedback to result in lasting improvements, it is hypothesized that the child must still attempt to recode the word in relation to the print—in this case, the pronunciation provided by the teacher may simply act as a catalyst to jumpstart future recoding opportunities.

The findings reported here are not in complete agreement with other studies in the self-teaching literature (e.g., Landi et al., 2006; Nation, Angell, & Castles, 2007; Ricketts, Bishop, Pimperton, & Nation, 2011; Wang et al., 2011, Experiment 1); however, there are several differences between this study and those reported previously that merit consideration. The first involves whether the dependent variable is the total number of words that can be *read correctly* (online performance) or the number of words *learned*. During the present experiment, the same children were substantially more successful when presented with words in context compared with in isolation. However, because this benefit was apparent on the first trial, the rate of learning did not differ between the context and the isolation conditions. Therefore, the condition that results in superior reading is a matter of opinion. On the one hand, it could be argued that children *learned* just as many words in isolation as they did in context. On the other hand, it could also be said that presenting words in contexts allows children to experience *greater success* while reading from the first session onward. Both arguments are equally valid and supported by the data.

Second, the role of context is most influential when reading is difficult. Therefore, the nature of the materials is an important factor with regards to context effects. As noted in the introduction, several studies in this area have used very small training sets (e.g., eight items) and words that can be easily decoded (e.g., CVCe words such as "yate"). When decoding is relatively simple it is unlikely that children would need the extra support provided by context. It is hypothesized that using a much larger word set—with words that varied greatly in letter length and morphological complexity—made the present study more sensitive to change, and allowed the advantages of context to be observed.

A final consideration is whether spelling or reading is the variable of interest. The current experiment focused on self-teaching in relation to word reading rather than spelling. However, there could be any number of different mappings between the underlying orthographic representation and the way a word is ultimately read. The child could read the word correctly and have a high quality representation; conversely, the child could read the word correctly and have a low quality or incomplete orthographic representation (Martin-Chang, Ouellette, & Madden, 2014). Therefore, the current findings cannot comment on whether the benefits associated with contextual reading would generalize to spelling. In fact, there is reason to believe that they may not. For example, Kyte and Johnson (2006) found that reading accuracy during their learning phase was not equally correlated with all of the tasks that comprised the orthographic learning task. Rather, it was most closely linked to the reading accuracy posttest measure and least linked with the orthographic choice task. Given the wealth of evidence showing that context does not aid (nor hinder) orthographic learning, it could be possible that the two factors involved in self-teaching (decoding and whole-word phonology) have differential effects on reading and spelling development.

Grapheme-to-phoneme decoding may have a more profound impact on spelling development (suggesting higher quality orthographic representations in memory), while matching whole-word orthography to phonology might suffice for reading accuracy but not precise spelling (suggesting representations that may not be fully specified). It is also important to consider that a child could conceivably read a word incorrectly (e.g., read the word "chaos" as "chase" or "city" as "kitty") and still have formed a stable orthographic representation of that word in memory, albeit one that is associated with an incorrect pronunciation.

Limitations and Future Directions

This study offers a different perspective on self-teaching because it observed how children learn to read a large bank of real words over multiple exposures. In many respects the materials were representative of those children encounter when they read independently, specifically, they contained many words of varying difficulty to be learned. However, using real words made it difficult to determine the proportion of items that were self-taught during the first session compared with those that were previously known. Therefore, it would be fruitful to replicate the procedures discussed here with words that were known verbally, but not in writing, to determine if the same patterns hold. In addition, the ability to arrive at a correct pronunciation without a high quality representation to support reading is partially driven by the properties of the words themselves. Therefore, examining the interaction between context and word regularity may be insightful. Future studies should consider using a cross-classified analytic approach (see Kim, Petscher, Foorman, & Zhou, 2010) to examine the contribution of both participant characteristics and word features simultaneously.

A second limitation involves the shared reading paradigm used during this study. Shared reading, while offering a high degree of experimental control, does not mimic the independent reading that generally happens when children self-teach. This manner of presentation might have felt awkward for good readers, who could have read the whole text alone. It may also have maximized the effects of context for poor readers, who might have struggled to read the surrounding text independently. Therefore, this study should be replicated when children are reading independently.

Conclusions

Orthographic learning has been attributed with providing readers with fast, accurate, and long lasting access to written words; in short, self-teaching is the mechanism that has been attributed with improving reading fluency. The results from the present experiment showed that children learned to read a number of words without feedback in both context and in isolation, but that improvement during training was more pronounced (and equally lasting) when feedback was provided. The data also showed that children could read substantially more words accurately when they were presented with words in context compared with in isolation. The most advantageous of all the conditions was when children were given the benefits of both context and feedback.

The conclusions drawn from the present study suggest that generating (with the help of context) or being given a word's pronunciation (via feedback) aids in accurate decoding on future word encounters. It would seem that children use all available resources, including feedback and context, to amalgamate the orthographic, phonological, and semantic properties of words as they are learning to read.

References

Barbetta, P. M., Heward, W. L., Bradley, D. M., & Miller, A. D. (1994). Effects of immediate and delayed error correction on the acquisition and maintenance of sight words by students with developmental disabilities. *Journal of Applied Behavior Analysis, 27,* 177–178. http://dx.doi.org/10.1901/jaba.1994.27-177

Bowey, J. A., & Muller, D. (2005). Phonological recoding and rapid orthographic learning in third-graders' silent reading: A critical test of the self-teaching hypothesis. *Journal of Experimental Child Psychology, 92,* 203–219. http://dx.doi.org/10.1016/j.jecp.2005.06.005

Byrne, B., Wadsworth, S., Boehme, K., Talk, A. C., Coventry, W., Olson, R., . . . Samuelsson, S. (2013). Multivariate genetic analysis of learning and early reading development. *Scientific Studies of Reading, 17,* 224–242. http://dx.doi.org/10.1080/10888438.2011.654298

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2 ed.). Hillsdale, NJ: Erlbaum.

Cunningham, A. E. (2006). Accounting for children's orthographic learning while reading text: Do children self-teach? *Journal of Experimental Child Psychology, 95,* 56–77. http://dx.doi.org/10.1016/j.jecp.2006.03.008

Cunningham, A. E., Perry, K. E., Stanovich, K. E., & Share, D. L. (2002). Orthographic learning during reading: Examining the role of self-teaching. *Journal of Experimental Child Psychology, 82,* 185–199. http://dx.doi.org/10.1016/S0022-0965(02)00008-5

Ehri, L. (2014). Orthographic mapping and the acquisition of sight word reading, spelling memory and vocabulary learning. *Scientific Studies of Reading, 18,* 5–21. http://dx.doi.org/10.1080/10888438.2013.819356

Ehrlich, S., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning & Verbal Behavior, 20,* 641–655. http://dx.doi.org/10.1016/S0022-5371(81)90220-6

Elbro, C., de Jong, P., Houter, D., & Nielsen, A. (2012). From spelling pronunciation to lexical access: A second step in word decoding. *Scientific Studies of Reading, 16,* 341–359. http://dx.doi.org/10.1080/10888438.2011.568556

Elgort, I., Perfetti, C. A., Rickles, B., & Stafura, J. Z. (2015). Contextual learning of L2 word meanings: Second language proficiency modulates behavioral and event-related brain potential (ERP) indicators of learning. *Language, Cognition and Neuroscience, 30,* 506–528.

Evans, M. A., Barraball, L., & Eberle, T. (1998). Parental responses to miscues during child-to-parent book reading. *Journal of Applied Developmental Psychology, 19,* 67–84. http://dx.doi.org/10.1016/S0193-3973(99)80028-8

Field, A. P. (2009). *Discovering statistics using SPSS* (3rd ed.). London, England: SAGE.

Gough, P. B., Alford, J. A., Jr., & Holley-Wilcox, P. (1981). Words and contexts. In O. J. L. Tzeng & H. Singer (Eds.), *Perceptions of print* (pp. 85–102). Hillsdale, NJ: Erlbaum Associates.

Kim, Y. S., Petscher, Y., Foorman, B., & Zhou, C. (2010). The contributions of phonological awareness and letter-name knowledge to letter sound acquisition—A cross-classified multilevel model approach. *Journal of Educational Psychology, 102,* 313–326. http://dx.doi.org/10.1037/a0018449

Kyte, C. S., & Johnson, C. J. (2006). The role of phonological recoding in orthographic learning. *Journal of Experimental Child Psychology, 93,* 166–185. http://dx.doi.org/10.1016/j.jecp.2005.09.003

Landi, N. (2013). Learning to read words: Understanding the relationship between reading ability, lexical quality, and reading context. In M. A. Britt, S. Goldman, & J. F. Rouet (Eds.), *Reading—From words to multiple texts* (pp. 17–33). New York, NY: Routledge.

Landi, N., Perfetti, C. A., Bolger, D. J., Dunlap, S., & Foorman, B. R. (2006). The role of discourse context in developing word form representations: A paradoxical relation between reading and learning. *Journal of Experimental Child Psychology, 94,* 114–133. http://dx.doi.org/10.1016/j.jecp.2005.12.004

Martin-Chang, S. L., & Levy, B. A. (2005). Fluency transfer: Differential gains in reading speed and accuracy following isolated word and context training. *Reading and Writing, 18,* 343–376. http://dx.doi.org/10.1007/s11145-005-0668-x

Martin-Chang, S. L., & Levy, B. A. (2006). Word reading fluency: A transfer appropriate processing account of fluency transfer. *Reading and Writing, 19,* 517–542. http://dx.doi.org/10.1007/s11145-006-9007-0

Martin-Chang, S. L., Levy, B. A., & O'Neil, S. (2007). Word acquisition, retention, and transfer: Findings from contextual and isolated word training. *Journal of Experimental Child Psychology, 96,* 37–56. http://dx.doi.org/10.1016/j.jecp.2006.08.004

Martin-Chang, S. L., Ouellette, G., & Madden, M. (2014). Does poor spelling equate to slow reading? The relationship between reading, spelling, and orthographic quality. *Reading and Writing, 27,* 1485–1505. http://dx.doi.org/10.1007/s11145-014-9502-7

Nagy, W., Anderson, R., & Herman, P. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal, 24,* 237–270. http://dx.doi.org/10.3102/00028312024002237

Nation, K., Angell, P., & Castles, A. (2007). Orthographic learning via self-teaching in children learning to read English: Effects of exposure, durability, and context. *Journal of Experimental Child Psychology, 96,* 71–84. http://dx.doi.org/10.1016/j.jecp.2006.06.004

Nation, K., & Snowling, M. J. (1998). Individual differences in contextual facilitation: Evidence from dyslexia and poor reading comprehension. *Child Development, 69,* 996–1011. http://dx.doi.org/10.1111/j.1467-8624.1998.tb06157.x

Ouellette, G. (2010). Orthographic learning in learning to spell: The roles of semantics and type of practice. *Journal of Experimental Child Psychology, 107,* 50–58. http://dx.doi.org/10.1016/j.jecp.2010.04.009

Ouellette, G., & Fraser, J. R. (2009). What exactly is a yait anyway: The role of semantics in orthographic learning. *Journal of Experimental Child Psychology, 104,* 239–251. http://dx.doi.org/10.1016/j.jecp.2009.05.001

Ricketts, J., Bishop, D., Pimperton, H., & Nation, K. (2011). The role of self-teaching in learning orthographic and semantic aspects of new words. *Scientific Studies of Reading, 15,* 47–70. http://dx.doi.org/10.1080/10888438.2011.536129

Roth, S. F., & Perfetti, C. A. (1980). A framework for reading, language comprehension, and language disability. *Topics in Language Disorders, 1,* 15–28. http://dx.doi.org/10.1097/00011363-198012000-00004

Samuels, S. J. (1967). Attentional process in reading: The effect of pictures on the acquisition of reading responses. *Journal of Educational Psychology, 58,* 337–342. http://dx.doi.org/10.1037/h0020045

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition, 55,* 151–218. http://dx.doi.org/10.1016/0010-0277(94)00645-2

Share, D. L. (1999). Phonological recoding and orthographic learning: A direct test of the self-teaching hypothesis. *Journal of Experimental Child Psychology, 72,* 95–129. http://dx.doi.org/10.1006/jecp.1998.2481

Share, D. L. (2004). Orthographic learning at a glance: On the time course and developmental onset of self-teaching. *Journal of Experimental Child Psychology, 87,* 267–298. http://dx.doi.org/10.1016/j.jecp.2004.01.001

Snowling, M. J., & Hulme, C. (2011). Evidence-based interventions for reading and language difficulties: Creating a virtuous circle. *British Journal of Educational Psychology, 81,* 1–23. http://dx.doi.org/10.1111/j.2044-8279.2010.02014.x

Stanovich, K. E., Nathan, R. G., West, R. F., & Vala-Rossi, M. (1985). Children's word recognition in context; spreading activation, expectancy, and modularity. *Child Development, 56,* 1418–1428. http://dx.doi.org/10.2307/1130461

Tunmer, W. E., & Chapman, J. W. (1995). Context use in early reading development: Premature exclusion of a source of individual differences? *Issues in Education, 1,* 97–100.

Tunmer, W. E., & Chapman, J. W. (2012). Does set for variability mediate the influence of vocabulary knowledge on the development of word recognition skills. *Scientific Studies of Reading, 16,* 122–140. http://dx.doi.org/10.1080/10888438.2010.542527

Wang, H.-C., Castles, A., Nickels, L., & Nation, K. (2011). Context effects on orthographic learning of regular and irregular words. *Journal of Experimental Child Psychology, 109,* 39–57. http://dx.doi.org/10.1016/j.jecp.2010.11.005

Wilkinson, G. S. (1993). *The Wide Range Achievement Test.* Wilmington, DE: Wide Range.

Ziegler, J., Perry, C., & Zorzi, M. (2014). Modeling reading development through phonological recoding and self-teaching: Implications for dyslexia. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 369,* 1–9.

(*Appendices follow*)

Appendix A

Stimuli Lists and Word Frequency Counts Per Million Words

List 1	Frequency	List 2	Frequency	List 3	Frequency	List 4	Frequency
After	682.59	About	3631.49	Admired	3.67	Away	730.9
Ahead	198.33	Above	48.88	Announced	5.67	Awful	63.41
Animals	39.08	Again	792.71	Arms	59.8	Bald	9.73
Behind	187.86	Always	655.25	Around	736.73	Because	1071.02
Below	28.04	Answered	15.2	Asked	216.25	Belt	24.35
Beneath	11.63	Beating	22.75	Baby	509.37	Boots	19.16
Board	64.16	Before	794.14	Back	2009.16	Boys	224.16
Both	295.33	Began	32.51	Bath	31.12	Branch	10.08
Bounced	2.96	Bird	45.45	Bathroom	61.67	Breeze	8.04
Bridge	45.71	Blackbird	.82	Bathtub	6.1	Broke	105
Cars	45.63	Books	67.76	Bedroom	36.71	Called	340.02
Chuckled	.08	Bright	44.41	Best	404.37	Camera	57
Close	219.43	Child	157.65	Bowls	2.18	Cannon	8.71
Coming	527.02	Chirping	.98	Brand	13.96	Caught	93.94
Could	1629.59	Clever	27.27	Carried	20.12	Check	278.98
Crash	28.65	Confessed	5.96	Change	240.35	Climb	19.75
Cross	55.04	Cried	12.98	Clean	121.24	Couldn't	N/A
Deer	8.71	Demanded	2.76	Closely	9.18	Crack	32.84
Drove	28.86	Different	209.53	Continue	49.55	Crows	2.76
Elizabeth	N/A	Discovered	28.76	Crowded	8.94	Dead	448.98
Excited	48.61	Each	253.25	Declared	6.53	Decided	88.65
Fallen	16.92	Estimate	4.76	Door	292.06	Diving	6.29
Families	22.33	Ever	709.22	Enough	501.33	Down	1490.3
Fast	137.45	Every	549.16	Even	875.92	Edge	23.51
Father	554.49	Everything	654.88	Everyone	241.65	Feet	120.73
Filled	27.18	Forest	18.88	Family	354.25	Fell	73
Followed	34.1	Found	396	Felt	119.82	Firecrackers	.71
Forward	72.33	Freeze	32.16	First	840.57	Five	285.45
Foxes	1.16	Guess	453.98	Good	2610.14	Flapping	1.67
Girls	208.35	Heart	244.18	Grandma	N/A	Fluttered	.06
Glanced	.63	Here	4525.25	Hall	51.94	Foot	64.92
Going	2123.29	Home	774.33	Just	4749.14	Getting	484.69
Group	73.76	Huge	48.37	Kept	89.39	Giggling	4.1
Growled	.12	Imitate	1.8	Laughed	10.69	Grinned	.24
Hands	236.53	Kind	590.69	Lots	60.16	Halfway	13.29
Happy	333.2	Knew	368.96	Louder	10.1	Head	371.51
Hated	28.22	Know	5721.18	Made	561.29	Help	921.12
Held	42.45	Languages	4.1	Making	222.53	High	195
Hello	N/A	Leaving	141.39	Metal	19.45	Hurry	173.65
Hike	6.53	Listening	62.84	Mind	484.61	Looked	120.9
Inched	.02	Many	359.43	Modern	18.24	Make	1387.75
Included	7.49	Maple	3.24	More	1298.59	Mice	6.57
Jumped	21.14	Minute	377.49	Never	1362.55	Mayer	N/A
Laughing	52.29	Myself	342.55	Nothing	853.61	Moment	187.04
Little	1446.39	Name	641.86	Ordered	36.96	Much	973.25
Lived	66.04	Nearby	8.33	Papa	N/A	Mustn't	N/A
Long	675.16	Nimbly	.1	Picked	69.29	Nest	11.1
Moved	69.33	Nobody	266.65	Pitcher	3.24	Next	452.75
Named	69.88	Notice	59.25	Plastic	18.76	Only	1083.71
Need	1294.9	Often	57.35	Politely	1.71	Owls	2.12
Noises	7.16	Once	344.88	Pulling	27.14	Park	72.12
Nuts	53.51	Other	735.39	Quickly	56.49	Place	602.67
Onto	36.69	Professor	69.57	Screams	16.9	Quietly	12.33
Others	99.24	Realized	35.96	Seems	167.55	Ready	387.8
Popped	7.92	Right	4008.39	Shall	185.12	Realize	79.06
Possible	114.04	School	333.12	Shiny	7.8	Returned	24.76
Rabbits	6.43	Single	72.08	Should	1061.94	Scare	33.57
Replied	1.16	Size	46.14	Smiled	4.92	Screeching	2.55

(*Appendices continue*)

Appendix A (*continued*)

List 1	Frequency	List 2	Frequency	List 3	Frequency	List 4	Frequency
Riverbed	.43	Slid	1.84	Soapsuds	N/A	Seen	384.96
Rotten	17.47	Snickered	.02	Splashing	1.1	Shots	28.37
Seats	21.76	Some	1727.24	Spouts	.06	Sign	133.27
Seemed	54.25	Song	93.69	Started	187.57	Sitting	94.39
Shrill	.47	Sound	143.39	Stepped	12.86	Small	124.96
Slowly	25.08	Stopped	75.37	Stick	97.12	Softly	4.73
Something	1500.16	Strange	86.43	Stood	25.78	Sounded	18.86
Sounds	156.27	Study	49.04	Stop	707.27	Sticks	13.61
Stacey's	N/A	Suddenly	55.96	Sweater	13.8	Such	291.22
Start	340.1	Things	692.88	Thank	1115.24	Sure	1099.82
Stay	515.65	Those	753.02	Thick	13.98	Take	1891.04
Still	788.73	Thought	808.47	Though	181.94	Teacher	55.73
Story	220.78	Thousands	27.65	Towel	14.16	Teddy	N/A
Sudden	33.47	Town	247.92	Trouble	223.55	These	904
Swamps	.88	Tracks	16.75	Turned	105.65	Think	2691.39
Tested	10.53	Tried	186.84	Twisted	10.59	Tiptoe	.88
Their	655.16	Unknown	15.18	Waiting	211.12	Told	699.59
Today	433.8	Ventured	.47	Want	2759.18	Took	342.24
Together	383.39	Walk	215.86	Warm	52.14	Towards	27.43
Trail	19.2	Went	411.51	Warned	15.84	Tree	65
Troll	2.71	Whistle	15.45	Water	225.06	Trying	448.02
Unsure	1.02	Whose	62.49	Whimpered	.08	Underfoot	N/A
Very	1241.25	Wildly	1.92	Whined	.1	Under	261.92
Voice	86.16	Woods	29.06	Will	2123.65	Warning	31.96
Who's	N/A	Would	1767.88	Wooden	7.2	Whispered	2.02
Wondered	14.9	Written	44.06	World	455.22	Without	354.65
Wonderful	164	Young	243.18	Years	568.69	Yelled	6.14
Average	231.88		443.38		377.65		291.04

(*Appendices continue*)

Appendix B

Example of a Training Story "Mary's Race Day"

(*target words are those bolded and underlined in the story)

The **thick door** flung open and the children ran outside onto Barton Street. They were all looking to see the **shiny** new wagon with its **modern metal** frame. Mary's was the **best** wagon in **years**. The children **admired** its fancy **plastic** steering wheel and its **brand** new **wooden** basket. They thought it was the fastest wagon in the **world**. The children **laughed** and yelled **louder** as **more** wagons **crowded** onto Barton Street.

Mary could hear the children's **screams** from her **bedroom**. She didn't **want** to keep them **waiting**. She **quickly** rushed through the **hall** into the **bathroom** for a **bath**. When Mary finished the **bath,** her **splashing** had left **water** and **soapsuds** all over the **bathtub**. She **picked** up a **towel** and ran it over the **spouts** and **bowls**. She **stepped back** and **admired** the **shiny bathtub** and **spouts**.

"Finished," she **declared**.

She left the **bathroom** and went into her **bedroom**. She **smiled** while **pulling** on her **warm thick sweater**. Today was Mary's **first** wagon race. **Nothing** could **change** her wonderful mood. Mary had **never felt more** prepared for anything. She **stepped back** into the **hall**. Her **towel** was full of **soapsuds** so she **carried** it to the **plastic bowls.** Then she went downstairs. In the kitchen she saw that her father had **picked lots** of lemons. He **twisted** and **turned** them until he **made** a **baby pitcher** of lemonade. **Papa** thought the old fashioned lemonade was better than the new **modern** stuff from a can. Mary loved it. She often drank a whole **baby pitcher** herself. But today she **politely declared** that she was full and that she **should stop**. Mary was careful to **mind** her manners and **thank** her father. She knew she had the **best family**. Mary wouldn't trade them for the **world**. Soon **everyone** was set to go so they left the house and locked the front **door**.

On the way to the race, **Grandma asked** Mary if she **should continue** to stretch her **arms** and run **around** the block to **warm** up her legs. Mary **smiled**. Her **Grandma never ordered** to do things. She **just politely** suggested them.

"That **seems** like a **good** idea," Mary **announced**, "I think I **shall** try it."

But when Mary **started** down the street she wanted to **change** her **mind**. There were **lots** of people yelling for her and they were **louder** than she could have imagined.

"Oh no," Mary **whined**. As her **family** began going towards the **screams** and **crowded** streets Mary's eyes began to **water**.

"I'm Scared," she **whined**.

"I know this **seems** difficult, but I have **years** of experience," her **Papa warned** "You **will continue** on alone but we **shall** follow **closely** behind."

Mary **felt** like **pulling** on his **arms** and **making** him come with her but she knew that she was old **enough** to go alone.

Even though she was frightened Mary **kept** her head up as she **stood** by her wagon. Her heart **whimpered** with fear as they were **ordered** to get into their wagons. But she wasn't **waiting** long. Soon the gun fired and the race had **started**. Mary **made** a **clean** getaway and she was **quickly** in **first** place. But then Tom, another racer, scooted **closely** past her. Tom also had a **brand** new **metal** wagon. He was **making** this an exciting race. As the road **twisted** down the hill, Mary spotted **trouble**. There was a large **wooden stick** in the middle of the road but there was **nothing** she could do. Tom could not be **warned**. He **just kept** racing towards **trouble**. His wagon came to a sudden **stop** as it hit the **stick**. Tom fell hard, **splashing** into a puddle. Some of the other kids **laughed** as they passed him but Mary didn't **want** to leave him there. She **turned** her wagon **around** and **carried** him off the road. She gave him her **sweater** and helped him **clean** his cuts. Tom could not **thank** her **enough**.

"**Will** you be alright?" Mary **asked**.

"Yes," he **whimpered**, "thanks to you."

At the end of the day, **even though** Mary lost the race, her **good** deed was **announced** and **everyone stood** and clapped for her.

Received August 14, 2015
Revision received April 6, 2016
Accepted April 8, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 2, 245–256

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000134

The Role of the Updating Function in Solving Arithmetic Word Problems

Kanetaka Mori and Masahiko Okamoto
Osaka Prefecture University

We investigated how the updating function supports the integration process in solving arithmetic word problems. In Experiment 1, we measured reading time, that is, translation and integration times, when undergraduate and graduate students ($n = 78$) were asked to solve 2 types of problems: those containing only necessary information and those containing extraneous information. The results indicated that participants required more integration time to solve extraneous-information problems than necessary-information problems. However, the higher the updating, the smaller the increment of integration time. In Experiment 2, we investigated whether different problem models were provided by undergraduate and graduate students ($n = 73$) with different updating functions. Participants executed a lexical-decision task immediately following an integration process. The lexical-decision task comprised 3 conditions: necessary-information word, extraneous-information word, and novel word conditions. The RTs for both necessary- and extraneous-information word conditions were faster than that for the novel word condition. The facilitation amount in an extraneous-information word became weaker as the problem solver's updating function increased. These results suggest that individuals with a high updating function provide a problem model that maintains only task-relevant information, while those with less-effective updating use an approach that also considers extraneous information. These 2 experiments indicate that updating is an important contributor to the integration process and different updating abilities result in different problem models.

Keywords: arithmetic word problems, integration, updating, problem model

Understanding cognitive processes underlying arithmetic word problems enables us to develop better instructional designs. Recently, researchers have investigated the relationship between working memory and arithmetic or mathematics. Arithmetic word problem solving requires active manipulation of relevant information. Working memory is used as a mental space for all cognitive activities including word problem solving. According to Baddeley's (1986) three-component model, working memory involves two subsystems: a phonological loop, which stores verbal materials, and a visuospatial sketchpad, which houses visual information. These two components are coordinated by the central executive, a supervisory system concerned with attention. Several studies have revealed the importance of the central executive for arithmetic word problem solving (Andersson, 2007; Fuchs et al., 2010; Lee, Ng, Ng, & Lim, 2004; Swanson, 2006; Swanson & Sachse-Lee, 2001). Lee et al. (2004) investigated the relationships among working memory, literacy, performance IQ, and arithmetic word problem performance. They revealed that the phonological loop and visuospatial sketchpad contribute to arithmetic word problem performance via literacy and performance IQ, respectively. They also found that the central executive contributes both directly and indirectly to arithmetic word problem performance. Arithmetic word problem solving requires problem solvers to not only maintain incoming information using the phonological loop and visuospatial sketchpad but also control the information through the central executive.

Although the central executive contributes to arithmetic word problem solving, we do not yet understand how this system supports problem solving. If we know more about the relationship between word problem solving and the central executive, we can gain implications for what cognitive process or function we should teach. In brief, we need to know how the central executive functions in solving arithmetic word problems.

Some researchers have suggested that the central executive has a range of functions (Baddeley, 1996; Miyake et al., 2000). Miyake et al. (2000) provided evidence that the central executive's functions include at least three unity and diversity functions, described as inhibiting, shifting, and updating. Inhibiting is the ability to inhibit dominant, automatic, and prepotent responses. Shifting is the ability to switch back and forth flexibly between tasks or mental sets. Updating is the ability to monitor incoming information for relevance to the task at hand and then appropriately update by replacing old, no longer relevant information with new, more relevant information. Furthermore, these functions overlap. That is, updating can include an inhibitory process. In fact, Miyake and Friedman (2012) reported that updating consists of a common executive function that substitutes for inhibition and the updating specific process. Although there is some discussion on this point, we follow Miyake et al.'s (2000) definition to relate our study to previous research.

This article was published Online First June 2, 2016.

Kanetaka Mori and Masahiko Okamoto, Department of Humanities and Social Sciences, Osaka Prefecture University.

This research was supported by grants from Research Fellowships of Japan Society for the Promotion of Science for Young Scientists.

Correspondence concerning this article should be addressed to Kanetaka Mori, Department of Humanities and Social Sciences, Okamoto Lab, Osaka Prefecture University, 1-1 Gakuen-cho, Naka-ku, Sakai, Osaka 599-8531, Japan. E-mail: k.mori.k@gmail.com

Agostino, Johnson, and Pascual-Leone (2010) investigated the relationships among the three executive functions, mental or M-capacity, and performance on arithmetic word problems. Results of structural equation modeling indicated that the updating function was the best predictor of accuracy in arithmetic word problems. Furthermore, Passolunghi and Pazzaglia (2005) demonstrated that children who achieve better mathematical performance show higher updating performance. These results indicate that the updating function relates to solving arithmetic word problems. Although investigation of the relationship between working memory and arithmetic word problem performance revealed that updating is important in solving arithmetic word problems, its role in solving arithmetic word problems remains unclear. To understand the arithmetic word problem-solving process more clearly, we need to determine the amount of updating required and the phase of problem solving influenced by updating.

Process of Arithmetic Word Problem Solving

Mayer's model (Mayer & Hegarty, 1996) and Polya's (1957) model suggest that the most difficult and important phase for individual problem solvers is to understand a problem. This comprehension phase can be divided into translation and integration phases (Kintsch, 1992; Mayer & Hegarty, 1996). In the translation phase, problem solvers need to translate each statement in a problem into a mental representation. In the integration phase, problem solvers must integrate each representation into a problem model. Problem solvers use numerical expressions to solve the problem, based on the problem model employed. The process that results in the formation of a particular model is important because the solution strategy depends on it.

Muth (1984, 1992) demonstrated that word problem-solving performance is influenced by the difficulty of integration. She assigned a problem-solving task in which the problem included extraneous information (not required to solve the problem). Her experiment showed that a child who could solve a problem with no extraneous information was likely to fail in solving an extraneous-information problem. Although extraneous information did not change the syntactic complexity or the expression of a problem, it increased demands related to integration of information. Muth (1984, 1992) also revealed that the integration process is important and extraneous information increases the difficulty in the process, but did not indicate why this effect occurs and what cognitive function relates to a successful integration process. Thus, we need to investigate the cognitive function underlying the integration process.

Hegarty, Mayer, and Green (1992) and Okamoto (1999) examined the nature of word problems that are challenging for children and adults, by looking at the reading times required for integration. They found that both children and adults require more time for integration when solving extraneous-information problems. These results suggest that extraneous-information problems and measurement of reading times are useful for examining the integration phase in solving word problems.

Relationship Between Integration and the Updating Function

An important assumption generated from research into the relationship between working memory and word problem solving is that the updating function may play a critical role in the integration process. In the integration process, one needs to continually integrate incoming information to form a problem model. When new information enters working memory, it changes the model and the old problem model is replaced by the newer one. This means that problem solvers might have to update the nature of their problem model sequentially when they are carrying out a word problem-solving task. Updating failure results in an inappropriate problem model and an incorrect answer. Therefore, we hypothesize that updating in integration is crucial for word problem solving.

Evidence to support this prediction comes from Kotsopoulos and Lee (2012), who examined the phase and nature of errors occurring in word problem solving. They video-recorded students talking aloud while completing homework. Their speech coding identified which of the four problem-solving phases appeared problematic for students and which singular executive function played the most important role in explaining their challenges. Updating challenges occurred when students were having difficulties evaluating information and appropriately editing it according to more relevant information in ways that would allow them to proceed to the next problem-solving phase. This research revealed that most errors occur in the integration process, and were caused by the updating function in integration. Integration might be a main process in solving word problems, and updating is a key cognitive function in this process. However, these results were obtained from subjects' verbal reports; this indicates that underlying cognitive processes are not clearly reflected. There is no direct evidence that their errors were due to the updating function, because Kotsopoulos and Lee did not objectively measure students' updating function. Thus, the relationship between the updating function and the integration process needs to be further investigated empirically.

Present Study

The purpose of this study was to reveal the role of the updating function in integration when solving an arithmetic word problem. Investigating this role would provide insights into how word problems are solved and what process we must carefully teach children who have lower updating.

If the updating function aids an efficient integration process, a problem solver who has higher updating function might be slightly influenced by the difficulty of integration, although extraneous information could increase the difficulty of integration (Muth, 1984, 1992). To explore the relationship between the updating function and the integration process, we measured reading times when undergraduate students were solving necessary- and extraneous-information problems.

The contribution of working memory to arithmetic word problems is not stable in elementary school students (Rasmussen & Bisanz, 2005; Meyer, Salimpoor, Wu, Geary, & Menon, 2010). Moreover, if problem solvers do not acquire language skills, they cannot solve arithmetic word problems, and so we cannot examine the contribution of working memory to arithmetic word problem solving. Because working memory and language skills are confounded in children, we used undergraduate students with stable working memory, sufficient language skills, and related problem-solving schema.

Our predictions were as follows: (a) An extraneous-information problem requires more integration time than a necessary-information problem; (b) the effect of extraneous information on integration time is smaller in higher updating problem solvers.

Experiment 1

Participants

Participants were 78 undergraduate students (42 males) in Japan with an average age of 19.54 years. Most participants were recruited from lectures of introduction to psychology in 2012 and 2015. They participated voluntarily, but were compensated with course credits or a book coupon for 500 yen. Some participants were recruited personally from acquaintances, who were not very familiar to the researcher. All participants were native Japanese speakers and were enrolled in the following courses: Social Science, Human Science, Engineering, Science, and Health Science. Math-related courses were Engineering and Science. Participants in other courses also had sufficient mathematical achievement, as demonstrated by their passing of the National Center Test for University Admissions in Japan. Sample problems of this test were reported in Wu (1993). Mathematics in this test includes vectors and functions.

Apparatus

Stimulus presentation and response recording were controlled by Matlab with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007) on an Apple iMac 21.5-inch display and an original USB response key box with three buttons.

Tasks and Procedures

Arithmetic word problem-solving task. Participants were asked to read a word problem and select a formula to solve it. Sentences were displayed one by one. Participants could manipulate the displayed sentence, moving backward or forward by pressing the left or right key on the response box. They were required to press the center key once they had comprehended the problem. After pressing this key, three expressions appeared, and participants selected the correct expression.

To analyze the comprehension process in arithmetic word problem solving, we calculated three types of reading times (see Figure 1). The time participants took to read a word problem sentence until expressions were displayed was the "whole reading" time. Time taken from onset of a problem to the end of the question statement's presentation was the "translation" time. Time taken from the question statement's initial presentation to the expressions' presentation was the "integration" time. This reflected the integration process in solving a word problem. Hegarty et al. (1992) and Okamoto (1999) measured the time from a question statement's presentation to selecting a numerical expression to solve a problem. This reflects not only the integration process but also the planning process. However, we must distinguish integration time from planning time. Thus, in this study, integration time was defined as the time from the onset of question presentation to the end of reading a problem. Planning time was defined as the reaction time (RT) needed to select the correct expression.

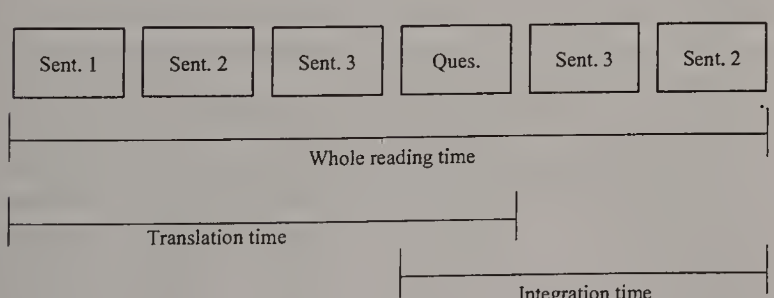


Figure 1. Example of three reading times. Sent. and Ques. each indicate a sentence in a problem. Ques. indicates an interrogative statement.

We prepared two types of arithmetic word problems: necessary-information and extraneous-information problems. Table 1 shows an example of each. A necessary-information problem consisted of three sentences—all needed to form an expression. The extraneous-information problem contained four sentences—one of which was unnecessary for forming an expression. We used 16 arithmetic word problems. Half of the problems were necessary-information problems and the others were extraneous-information problems. Four problems were used for practice trials.

Updating tasks. We used two updating tasks: (a) a letter memory task (adapted from Miyake et al., 2000) for the phonological domain, and (b) a visual *n*-back task (adapted from Agostino et al., 2010) for the visual domain.

The letter memory task measured phonological updating in working memory. In this task, letters were presented serially on a computer screen at a rate of 2,000 ms per letter. Participants were required to rehearse aloud the last four letters throughout by dropping the fifth letter and adding the most recent one. We recorded participants' verbal recall. The numbers of letters in the stimulus sets were 5, 7, 9, and 11. Each set was presented four times, and each participant completed 16 trials. If participants could accurately name aloud all letters in each set, the trial was classified as correct. The phonological updating score was the percentage of correct trials.

The visual *n*-back task measured visual updating of working memory. Three dots were presented on a computer screen. Participants were required to decide whether the current pattern was the same as the pattern *n* trials before. Their decisions were recorded by pressing a key that changed the presentation pattern. In each *n*-back task, there were 54 test trials following 12 practice trials. We used 20 matched test trials and 34 mismatched trials. We used the 1-back and 2-back tasks. Each score was calculated by the following formula: (proportion of correct match + proportion of correct mismatch).

General Procedures

Participants were asked to perform the arithmetic word problem task and then two updating tasks. The order of the phonological and visual updating tasks was counterbalanced across participants. Participants received instructions, which were displayed on the monitor, prior to undertaking each task, and they could ask the experimenter any question. The total experimental time was approximately 40 min.

Table 1
Examples of Necessary-Information Problems and Extraneous-Information Problems

Necessary-information problem	Extraneous-information problem
There are 5 pencils. There are 1.6 times more pens than pencils. How many pens are there?	There are 23 apples. There are 3 times more oranges than apples. <u>There are 8 times more grapes than apples.</u> How many oranges are there?

Note. The underlined sentence is extraneous information.

Results and Discussion

One participant was excluded from the analysis because of low accuracy (68%) in selecting a correct expression. Thus, data from 77 participants were analyzed. Whole reading, translation, integration, and planning times were calculated after excluding the error trials. Trials were excluded if an individual's whole reading time exceeded the mean $\pm$ 2 standard deviations (*SD*). For analysis of planning time, trials over the mean $\pm$ 2 *SD* for planning time were excluded. All reading times were standardized by each subject because there were individual differences in times (whole reading, translation, and integration times). Table 2 shows basic statistics for all measures. The phonological updating score was measured by a letter memory task, and the visual updating score was obtained using a visual 2-back task.

Accuracies for selecting a correct expression were analyzed using a paired sample t test for word problem type (necessary-information problem vs. extraneous-information problem). This analysis showed a significant difference in accuracy between word problem types, $t(76) = 2.98$, $p < .01$, which indicated that participants made more errors with extraneous-information problems than with necessary-information problems. Similarly, whole reading time was analyzed using a paired sample t test for word problem type. This analysis also showed a significant difference, $t(76) = -22.69$, $p < .001$, indicating that participants had more difficulty and needed more time to comprehend an extraneous-information problem than a necessary-information problem.

Individual differences of updating in solving word problems. To estimate the contribution of the updating function in solving word problems, we used a multiple regression analysis, with translation, integration, and planning times as dependent variables and problem type, phonological updating score, visual updating score, and these interactions as independent variables. Before the analysis, the phonological updating score and visual updating score were centered on the mean. Problem type was coded as a dummy variable, with 0 indicating a necessary-information problem and 1 an extraneous-information problem. Table 3 shows results of regression analysis for each dependent variable.

Analysis of translation time showed that phonological updating did not show any significant contribution. This result suggested that the phonological updating function was not important in translation. Furthermore, word problem type and interaction of visual updating and word problem type were significant, $t(146) = 21.92$, $p < .001$; $t(146 = -2.20$, $p < .05)$. These results indicated that although translation time was longer for extraneous-information problems than for necessary-information problems, higher visual updating reduced this tendency. The requirement for more translation time with extraneous-information problems was mostly due to the extraneous-information sentence. Higher visual updating solvers might be faster at encoding sentences than lower visual updating solvers. This effect was observed especially in extraneous-information problems because these problems included an additional sentence to be encoded.

In integration reading time, regression analysis revealed a significant contribution to word problem type, $t(146) = 12.63$, $p < .001$. This further revealed a significant contribution for the interaction of phonological updating and word problem type, $t(146) = -2.01$, $p < .05$, but not for phonological updating. These results indicated that in necessary problems, the phonological updating function did not contribute to integration, while in extraneous problems, the higher phonological updating function reduced integration time. These results support our hypothesis that the updating function aids an efficient integration process.

Analysis for planning time did not show any significant contribution. Even problem type was not significant. This indicated that planning time was the same for necessary-information and extraneous-information problems, suggesting that problem solvers completed the integration process before selecting an expression in this experiment.

In summary, in their integration, lower phonological updating solvers were more strongly affected by extraneous information than higher phonological updating solvers. The integration time difference might be caused by differences in their updating functions. Specifically, findings regarding integration time indicate that the integration process depends on the updating function. If the integration process differs by the updating function, the problem model that arises from the integration process can be different. However, the nature of the problem model resulting after the integration process was not clear in this experiment and requires further investigation.

According to Kintsch (1992), the integration process includes elaboration of the problem model. In this elaboration, one forms an appropriate problem model that activates only the information required to solve the problem. However, extraneous information is not activated in such a clear problem model. One possibility that can account for the difference between phonological updating abilities is that the updating function can help elaborate a problem to construct a clear problem model.

An alternative explanation is that working memory capacity (WMC) is also available. Problem solvers with high WMC could

Table 2
Basic Statistics for Whole Reading, Translation, Integration, and Planning Times; Accuracy in Selecting an Expression; and Updating Functions in Experiment 1

Variable	*M*	*SD*	Min	Max
Whole reading time	10,732.10	3,959.68	3,621.33	23,049.50
Translation time	8,949.21	3,771.67	2,588.80	23,049.50
Integration time	3,944.38	1,629.80	1,275.00	9,757.83
Planning time	1,781.65	573.97	804.00	4,382.50
Accuracy	.94	.07	.75	1.00
Phonological updating	.62	.25	.06	1.00
Visual updating	.84	.10	.46	1.00

Table 3
Standardized Partial Regression Coefficients of Regression Analysis for Translation, Integration, and Planning Times in Experiment 1

Independent variable	Translation time	Integration time	Planning time
Word problem type	1.75***	1.43***	.03
Phonological updating	−.06	.09	−.07
Visual updating	.10	−.08	.07
Phonological updating × Visual updating	.04	.02	.05
Phonological updating × Word problem type	.08	−.23*	−.12
Visual updating × Word problem type	−.18*	.16	.05
Three-way interaction	−.07	−.04	−.04
Adjusted R-squared	.76	.51	−.02
$F(7, 146)$	69.75***	23.80***	.58

Note. Asterisks indicate significance of the coefficients (* = .05. *** = .01).

easily solve arithmetic word problems compared with problem solvers who had low WMC (e.g., Rasmussen & Bisanz, 2005; Geary, 2004). If problem solvers with high WMC form more appropriate problem models than those with low WMC, the model provided by the high WMC can include all information linked to a problem. This problem model might use extraneous information to ensure high activation. Investigation into the specific nature of activation based on information included in a problem model suggests which explanation is appropriate. Therefore, we need to explore the level of activation of both necessary and extraneous information after integration.

Experiment 2

The purpose of the experiment was to reveal the nature of the problem model as a function of updating. We used a lexical-decision task for this purpose. Participants were required to decide whether a presented stimulus was a word or not. This type of decision is made more rapidly if the same word has been previously presented, because the stimulus word's representation is already activated. A lexical-decision task is useful for investigating the representation's activation in working memory.

We predicted response times (RTs) in the lexical-decision task. If problem solvers with high updating formed a clear problem model, RTs for necessary-information words could be faster than those for extraneous-information words. In lower updating solvers, representations of both extraneous- and necessary-information words could still be activated. This is because their problem model could not be updated and they were likely to form a problem model including extraneous information. Therefore, in comparison to a novel word, the lexical decision for these two words could be more rapid. The RTs for extraneous- and necessary-information words might not differ for lower updating solvers, while RTs for necessary-information words could be faster than those for extraneous-information words in higher updating solvers. Investigating RTs for these three types of words in a lexical-decision task would enable us to explore the nature of the problem model.

Participants

In Experiment 2, 73 undergraduate and graduate students in Japan (29 males) participated. Their average age was 19.59 years. Most participants were recruited from lectures of introduction to psychology in 2013 and 2015. Although they participated voluntarily, they received course credits or a book coupon worth 500 yen. Some participants were recruited personally from the researcher's acquaintances, who were not very familiar. All participants were native Japanese speakers and were enrolled in the following courses: Social Science, Human Science, Engineering, Science, and Health Science. Math-related courses were Engineering and Science. None had participated in Experiment 1.

Tasks and Procedures

Arithmetic word problem-solving task. To explore activation immediately after the integration process, we assigned the lexical-decision task between reading a problem and selecting a correct expression.

First, participants were instructed to read a problem until they comprehended it. In the first phase, each sentence of a problem was displayed individually so that participants were free to choose which sentence they wanted to read. After participants pressed a key that indicated they had understood the problem, they were given the lexical decision phase. In this phase, two stimuli (consisting of a word and a nonword) were presented horizontally on a computer screen. Participants were required to select a word from these two stimuli by pressing a key (the left key for the stimulus on the left and the center key for the stimulus on the right). After the lexical decision phase, two expressions were presented horizontally on the computer screen. One of the two expressions was the correct solution to the arithmetic word problem. Participants were required to select the correct expression by pressing a key.

A set of 24 arithmetic word problems was developed. Half of the problems used extraneous information that was selected in a pseudorandom manner. This selection was counterbalanced across participants. Extraneous-information problems had an assignment, two related sentences, and a question. The two sentences included an extraneous sentence, which was unnecessary for solving the problem, and a relevant sentence, which was necessary for solving it. A variable noun in each sentence was called the extraneous-information word or the necessary-information word for this study's purposes.

Although a typical lexical-decision task requires a decision about a word or a nonword in one presented stimulus, the lexical-decision task in this study required participants to select a word stimulus from two-letter strings. After the lexical-decision task,

selection of a correct expression from two expressions was immediately presented. We used this lexical-decision task to reduce cognitive load from task switching. In the lexical-decision task, the presented words had three conditions: (a) the necessary-information word condition, (b) the extraneous-information word condition, and (c) the novel word condition. Based on word frequency norms for Japanese (NTT database series; Amano & Kondo, 1999), 72 words written in Hiragana were collected. Half of the words had two characters and the other half had three. We prepared 24 sets of three words that corresponded to the three conditions. A set of three words had the same length of characters and similar frequency. By linking two-letter syllables not normally associated with each other (Umemoto, Morikawa, & Ibuki, 1955), 24 nonwords were created. Nonwords comprising three letters were formed by adding one letter to two-letter syllables. Furthermore, six words and four nonwords were prepared for four practice problems.

Updating tasks. A letter memory task and a visual *n*-back task were used to index the updating function. The visual *n*-back task was devised based on two main features from Experiment 1 (adapted from Friedman et al., 2008). The first feature was the presentation duration. In Experiment 1, the stimulus was displayed until participants pressed a key. To ensure more updating in Experiment 2, the stimulus was displayed for 1,000 ms, followed by a blank screen for 1,500 ms. Participants pressed a key indicating whether or not the stimulus was same as the one that had been displayed as the *n*-previous stimulus in that blank. The second feature was the presentation stimulus. One black square and nine white squares were presented on a computer screen. As in Experiment 1, the phonological updating score was the percentage of correct responses and the visual updating score was the 2-back task score.

Results and Discussion

Analysis of word problem-solving process in problems containing extraneous information. Whole reading, translation, integration, and planning times were analyzed after excluding error trials to select an expression. Trials were also excluded for analysis if whole reading time was over the mean ± 2 *SD* for each subject. For analysis of planning time, trials over the mean ± 2 *SD* for planning time were excluded. One participant was excluded because she made many errors in the lexical-decision task and struggled to select the correct expression (correct responses were 92% in the lexical-decision task and 92% in selecting an expression). Table 4 shows the basic statistics for all measures.

Accuracy in selecting correct expressions was analyzed using a paired sample *t* test for word problem type (necessary vs. extraneous). This analysis showed a significant difference between word problem types, $t(71) = 2.65$, $p < .01$. Furthermore, whole reading time was analyzed using a paired sample *t* test for word problem type. This analysis showed that the difference between whole reading times for word problem types was significant, $t(71) = -25.03$, $p < .001$. These results indicated that extraneous-information problems are more difficult than necessary-information problems for participants' comprehension and solution.

Same as that in Experiment 1, translation, integration, and planning times were analyzed with the multiple regression method. Table 5 shows results of regression analysis. Word problem type, phonological updating, visual updating, and these interactions were used as independent variables. The results showed a significant contribution of word problem type to translation time, $t(136) = 20.57$, $p < .001$. For integration reading time, the results revealed that the contribution of word problem type, $t(136) = 17.82$, $p < .001$ and interaction of phonological updating and word problem type, $t(136) = -2.17$, $p < .05$ were significant. These results indicated that lower phonological updating solvers were more influenced by extraneous information than higher phonological updating solvers.

Analysis of planning time showed that the contribution of phonological updating was significant, $t(136) = -3.58$, $p < .001$. The results indicated that the higher the phonological updating, the faster the selection of a correct expression. The results of the word problem-solving process in Experiment 2 were consistent with those in Experiment 1.

Nature of problem model and updating. We required participants to execute the lexical-decision task immediately after the integration process to reveal differences between problem models of high and low updating solvers. The lexical-decision task for extraneous problems was analyzed.

Table 4 shows that the accuracy for lexical decisions was very high. Trials over the mean ± 2 *SD* for RTs and error trials in the lexical-decision task and in selecting an expression were excluded. RTs for lexical decision tasks were also analyzed using multiple regression analysis. To analyze target word type, we created two dummy variables: a necessary condition and an extraneous condi-

Table 4
Basic Statistics for Reading Time in Each Phase, Accuracy for Selecting an Expression, Reaction Time, Accuracy for Lexical Decision, and Updating Functions in Experiment 2

Variable	*M*	*SD*	Min	Max
Whole reading time	10,054.06	3,311.47	3,511.25	21,838.50
Translation time	8,633.82	3,114.11	1,994.92	15,727.18
Integration time	3,220.31	1,581.40	941.75	11,158.83
Planning time	1,117.98	345.73	491.00	2,620.75
Accuracy	.98	.03	.83	1.00
RTs for lexical decision	863.71	214.14	513.75	1,929.25
Accuracy for lexical decision	1.00	.01	.92	1.00
Phonological updating	.65	.22	.19	1.00
Visual updating	.92	.11	.49	1.00

Table 5
Standardized Partial Regression Coefficients of Regression Analysis for Translation, Integration, and Planning Times in Experiment 2

Independent variable	Translation time	Integration time	Planning time
Word problem type	1.73***	1.66***	.26
Phonological updating	−.02	.10	−.40***
Visual updating	−.03	.01	.10
Phonological updating × Visual updating	.02	.04	−.02
Phonological updating × Word problem type	.04	−.20*	−.01
Visual updating × Word problem type	.09	−.02	−.04
Three-way interaction	−.04	−.08	−.04
Adjusted *R*-squared	.75	.69	.14
F(7, 136)	60.92***	46.07***	4.27***

Note. Asterisks indicate significance of the coefficients (* = .05. *** = .01).

tion. Target word type was coded as follows: Necessary-information words were coded as 1 in the necessary condition and 0 in the extraneous condition. Extraneous-information words were coded as 0 in the necessary condition and 1 in the extraneous condition. Novel words were coded as 0 in the necessary condition and 0 in the extraneous condition. Two dummy variables, phonological updating, visual updating, and these interactions were used as independent variables. Table 6 shows results of this regression analysis.

The results showed significant contributions of phonological updating, $t(204) = -4.95, p < .001$, visual updating, $t(204) = 2.53, p < .05$, necessary condition, $t(204) = -10.34, p < .001$, and extraneous condition, $t(204) = -7.03, p < .001$. The results indicated that the higher the phonological updating, the shorter the RTs for lexical decisions. On the other hand, the higher the visual updating, the longer the RTs for lexical decisions. The results of target word type showed that RTs for necessary-information words were faster than those for extraneous-information and novel words, and the RTs for extraneous-information words were faster than those for novel words. Furthermore, interaction of phonological updating and extraneous condition was significant, $t(204) = -2.08, p < .05$. This indicated that the influence of phonological updating for RTs for extraneous-information words was weaker than that for novel and necessary-information words.

These results showed that both necessary- and extraneous-information words were strongly activated for lower phonological updating solvers. Conversely, the priming effect of necessary information was stronger than that of extraneous information for higher phonological updating solvers. Necessary information was activated more strongly than extraneous information in a problem model. High phonological updating solvers formed a clear problem model by updating problem models during integration. Those with lower phonological updating, even after integration, had equal activation of necessary information and extraneous information.

In summary, results of the lexical-decision task revealed that the nature of the problem model depends on individual differences related to phonological updating. Problem solvers with high phonological updating constructed a problem model that included task-relevant information only. Thus, the integration process in word problem solving depends on one's updating function. Problem solvers with low phonological updating might construct their problem model with all relevant and extraneous information. A problem solver's updating function is one of the most important cognitive factors in constructing a problem model during the integration process.

General Discussion

Constructing a Problem Model and the Role of the Updating Function

The purpose of this study was to reveal the relationship between the integration process and the updating function in solving arithmetic word problems. In two experiments, we found that integra-

Table 6
Standardized Partial Regression Coefficients of Regression Analysis for RTs in Lexical Decision Tasks in Experiment 2

Main effect		Two-way interaction		Three-way interaction	
Necessary condition	−1.34***	Phonological updating × Visual updating	.01	Phonological updating × Visual updating × Necessary condition	.01
Extraneous condition	−.91***	Phonological updating × Necessary condition	.22	Phonological updating × Visual updating × Extraneous condition	.06
Phonological updating	−.45***	Phonological updating × Extraneous condition	.27*		
Visual updating	.25*	Visual updating × Necessary condition	−.12		
		Visual updating × Extraneous condition	−.09	Adjusted *R*-squared = .40 *F*(11, 204) = 14.19***	

Note. Asterisks indicate significance of the coefficients (* = .05. *** = .01).

tion time for an extraneous problem was longer than when necessary information alone was given. Additionally, the effect of extraneous information on integration was stronger in a low phonological updating solver than in a high phonological updating solver. These results suggested that low updating solvers struggled to form a problem model, especially for problems involving extraneous information. The results support our hypothesis that updating is important in the integration process in solving arithmetic word problems.

The results of Experiment 2 support the assumption that differences in the updating function cause differences in the problem model. In the lexical-decision task, the higher the phonological updating, the faster the decisions. However, this effect was weaker for the extraneous-information word condition. This indicated that activation of necessary information was higher than that of extraneous information in higher phonological updating solvers. However, in lower phonological updating solvers, decisions for necessary- and extraneous-information words might be equally facilitated. This indicated the same activation of necessary and extraneous information. Note that both types of information were highly activated. High phonological updating solvers formed a problem model that included only necessary information, while low phonological updating solvers created a problem model that also included extraneous information. These problem models might influence planning time. Results of planning time showed that the higher the phonological updating, the shorter the planning time. Higher phonological updating solvers could form a correct expression based on a clear problem model, while for lower phonological updating solvers, planning should be based on an appropriate problem model that included extraneous information.

The results of Experiment 2 are consistent with a study by Passolunghi and Pazzaglia (2004) who investigated recall error reported by low and high updating participants after solving word problems. Their results indicated that the correct recall of necessary information was higher in the high updating than in the low updating group. High updating solvers maintained necessary information in working memory. According to our findings, these results might occur because of the activation of the information included in a problem model after the integration process.

The findings in our study revealed that the integration process in word problem solving depends on the updating function. This is because problem solvers update a problem model to form an appropriate problem model during comprehension.

Updating or Working Memory Capacity?

It is important to consider whether the facilitation of integration was an effect of updating or simply WMC. It may be the case that an individual who has a large WMC can perform updating tasks efficiently and reduce integration time because he or she can store all information in working memory (Daneman & Carpenter, 1980). If this is the case, such individuals are expected to integrate all information into a problem model easily.

However, this possibility was rejected because of our results from Experiment 2. As for results of the lexical-decision task, in lower phonological updating solvers, the facilitation effect was equal for necessary- and extraneous-information words. This was inconsistent with the account of simple WMC because the capacity of low phonological updating individuals was enough to store all information. High updating solvers updated their problem model. These results were more likely caused by the updating function rather than WMC.

On the other hand, recently, WMC seemed to exhibit our ability to control our attention (e.g., Engle, Tuholski, Laughlin, & Conway, 1999; Kane, Bleckley, Conway, & Engle, 2001). From this viewpoint, the present study's results were caused by WMC as general attentional control rather than the updating function. Indeed, the updating function has been shown to be highly correlated with WMC as measured by the operation span task (Miyake et al., 2000; St Clair-Thompson & Gathercole, 2006). Although the updating function is an underlying process of WMC, there is controversy over the relationship between updating function and WMC.

However, our results revealed that constructing an appropriate problem model does not mean the ability to maintain all information as highly activated, but rather, the ability to activate only the necessary information. This view of the updating function seems to account better for our results.

Origin of Updating

There are some studies focusing on the biological mechanism of updating. Dahlin, Neely, Larsson, Bäckman, and Nyberg (2008) demonstrated the way in which two updating tasks (a letter memory task and a 3-back task) indicated overlapping activation of the prefrontal cortex (PFC) and striatum. She also showed that activation increases in the striatum after 5 weeks of updating task training. This suggests that the updating function is related to the striatum. O'Reilly's prefrontal cortex basal ganglia working memory model also attempted to demonstrate that updating requires the striatum, which is part of the basal ganglia (O'Reilly, 2006; O'Reilly & Frank, 2006). According to this model, the PFC maintains task-relevant information and the basal ganglia offer selective gating of some PFC regions. If the state of the striatum is "No-Go," updating does not occur and information in working memory is maintained. If the striatum state is "Go," information in working memory is updated. Thus, task-relevant information is stored and updated in working memory.

This model could account for the results of the present study. In Experiment 2, high updating caused differential activation for necessary and extraneous information. Conversely, low updating led to similar activation with necessary and extraneous information. According to the prefrontal cortex basal ganglia working memory model, in a high updating individual, task-relevant information is maintained and extraneous information is excluded from working memory via a selective gating mechanism. In contrast, problem solvers with low updating may have deficits in this gating mechanism, such that extraneous information might be retained in working memory. Increase of integration time in low updating solvers could reflect that they attempt to update more frequently. These accounts could link studies for word problem solving with associated biological mechanisms and computational modeling. There is a need for further consideration of these topics.

Updating and Modality

Domains in the updating function should also be considered. In the present study, only phonological updating showed a relation-

ship with integration time. Visual updating did not indicate such a relationship. The reason for this could be that processing word problems requires verbal encoding, which depends on the phonological loop. Consequently, verbal information was encoded during integration, and only updating in the phonological domain showed a relationship with integration. According to this interpretation, updating of the visuospatial sketchpad did not play an important role in solving word problems in this study. However, the visuospatial sketchpad did, in some cases, contribute more to the word problem than the phonological loop. The contribution to word problems changed with age over time, even when problems were the same (Meyer et al., 2010). Second-grade students showed contribution of the phonological loop and central executive, while third-grade students demonstrated contribution of the visuospatial sketchpad. Processing word problems depended on the mental model of third-grade students who were skilled in translating words into mental representations, which requires the visuospatial sketchpad. These findings lead to a possibility that adults are more likely to encode word problems as mental models. Nevertheless, the present study did not show the importance of visual updating. Extraneous information included in the present study was of a verbal nature. Problem solvers could make decisions about relevance based on the verbal information given. The relationship between phonological updating and integration was strongly demonstrated. This suggests the possibility that problems demanding encoding of visual information could show a contribution of visual updating. Such problems (e.g., geometric problems) should be investigated to further understand the relationship between the updating function and integration.

Educational Implications

Our findings suggest that the less updating function causes unsuccessful integration. This is an explanation about unsuccessful integration from cognitive process perspective. Our findings suggest the possibility that children with lower updating have a problem with the integration process. These children might improve their performance in solving word problems if we could train their updating function with working memory training. Although working memory training has recently increased and its significance has been reported (Au et al., 2015), the training's effectiveness in solving arithmetic word problems needs further investigation. Our findings identified a process in which updating is important. This would help to develop such training.

Our findings also suggest that children with lower updating function will struggle with integration due to the need for updating their problem model. In turn, this suggests a way in which problem solvers with lower updating may be helped. Their performance could be improved when problem model updating is not required, that is, if problem solvers know how to construct a problem model in advance, they might be able to construct a problem model that includes the relevant information. Indeed, it was reported that problems that use a question as the first sentence are easier (Robinson & Hayes, 1978). In this case, problem solvers could activate their problem schema before reading the problem and then they could integrate the information based on the schema. This would reduce the updating load. Therefore, our findings emphasize that when problem solvers have difficulties with their updating, it is important to design instruction to reduce updating.

References

Agostino, A., Johnson, J., & Pascual-Leone, J. (2010). Executive functions underlying multiplicative reasoning: Problem type matters. *Journal of Experimental Child Psychology, 105,* 286–305. http://dx.doi.org/10.1016/j.jecp.2009.09.006

Amano, S., & Kondo, T. (1999). *Lexical properties of Japanese: Vol. 1. Word familiarity* (NTT database series). Tokyo, Japan: Sanseido.

Andersson, U. (2007). The contribution of working memory to children's mathematical word problem solving. *Applied Cognitive Psychology, 21,* 1201–1216. http://dx.doi.org/10.1002/acp.1317

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: A meta-analysis. *Psychonomic Bulletin & Review, 22,* 366–377. http://dx.doi.org/10.3758/s13423-014-0699-x

Baddeley, A. D. (1986). *Working memory.* New York, NY: Oxford University Press.

Baddeley, A. (1996). Exploring the central executive. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 49,* 5–28. http://dx.doi.org/10.1080/713755608

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436. http://dx.doi.org/10.1163/156856897X00357

Dahlin, E., Neely, A. S., Larsson, A., Bäckman, L., & Nyberg, L. (2008). Transfer of learning after updating training mediated by the striatum. *Science, 320,* 1510–1512. http://dx.doi.org/10.1126/science.1155466

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19,* 450–466. http://dx.doi.org/10.1016/S0022-5371(80)90312-6

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128,* 309–331. http://dx.doi.org/10.1037/0096-3445.128.3.309

Friedman, N. P., Miyake, A., Young, S. E., Defries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General, 137,* 201–225. http://dx.doi.org/10.1037/0096-3445.137.2.201

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Hamlett, C. L., Seethaler, P. M., . . . Schatschneider, C. (2010). Do different types of school mathematics development depend on different constellations of numerical versus general cognitive abilities? *Developmental Psychology, 46,* 1731–1746. http://dx.doi.org/10.1037/a0020662

Geary, D. C. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities, 37,* 4–15. http://dx.doi.org/10.1177/00222194040370010201

Hegarty, M., Mayer, R. E., & Green, C. E. (1992). Comprehension of arithmetic word problems: Evidence from students' eye fixations. *Journal of Educational Psychology, 84,* 76–84. http://dx.doi.org/10.1037/0022-0663.84.1.76

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General, 130,* 169–183. http://dx.doi.org/10.1037/0096-3445.130.2.169

Kintsch, W. (1992). A cognitive architecture for comprehension. In H. L. Pick, Jr., P. W. van den Broek, & D. C. Knill (Eds.), *Cognition: Conceptual and methodological issues* (pp. 143–163). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10564-006

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception, 36,* 1–16.

Kotsopoulos, D., & Lee, J. (2012). A naturalistic study of executive function and mathematical problem-solving. *The Journal of Mathematical Behavior, 31,* 196–208. http://dx.doi.org/10.1016/j.jmathb.2011.12.005

Lee, K., Ng, S. F., Ng, E. L., & Lim, Z. Y. (2004). Working memory and literacy as predictors of performance on algebraic word problems. *Jour-*

nal of Experimental Child Psychology, 89, 140–158. http://dx.doi.org/10.1016/j.jecp.2004.07.001

Mayer, R. E., & Hegarty, M. (1996). The process of understanding mathematical problems. In R. J. Sternberg & T. Ben-Zeev (Eds.), *The nature of mathematical thinking* (pp. 29–53). Hillsdale, NJ: Erlbaum.

Meyer, M. L., Salimpoor, V. N., Wu, S. S., Geary, D. C., & Menon, V. (2010). Differential contribution of specific working memory components to mathematics achievement in 2nd and 3rd graders. *Learning and Individual Differences, 20,* 101–109. http://dx.doi.org/10.1016/j.lindif.2009.08.004

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science, 21,* 8–14. http://dx.doi.org/10.1177/0963721411429458

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41,* 49–100. http://dx.doi.org/10.1006/cogp.1999.0734

Muth, K. D. (1984). Solving arithmetic word problems: Role of reading and computational skills. *Journal of Educational Psychology, 76,* 205–210. http://dx.doi.org/10.1037/0022-0663.76.2.205

Muth, K. D. (1992). Extraneous information and extra steps in arithmetic word problems. *Contemporary Educational Psychology, 17,* 278–285. http://dx.doi.org/10.1016/0361-476X(92)90066-8

Okamoto, M. (1999). *The study of metacognition in arithmetic word problem solving*. Tokyo, Japan: Kazamashobou.

O'Reilly, R. C. (2006). Biologically based computational models of high-level cognition. *Science, 314,* 91–94. http://dx.doi.org/10.1126/science.1127242

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation, 18,* 283–328. http://dx.doi.org/10.1162/089976606775093909

Passolunghi, M. C., & Pazzaglia, F. (2004). Individual differences in memory updating in relation to arithmetic problem solving. *Learning and Individual Differences, 14,* 219–230. http://dx.doi.org/10.1016/j.lindif.2004.03.001

Passolunghi, M. C., & Pazzaglia, F. (2005). A comparison of updating processes in children good or poor in arithmetic word problem-solving. *Learning and Individual Differences, 15,* 257–269. http://dx.doi.org/10.1016/j.lindif.2005.03.001

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442. http://dx.doi.org/10.1163/156856897X00366

Polya, G. (1957). *How to solve it: A new aspect of mathematical method* (2nd ed.). Princeton, NJ: Princeton University Press.

Rasmussen, C., & Bisanz, J. (2005). Representation and working memory in early arithmetic. *Journal of Experimental Child Psychology, 91,* 137–157. http://dx.doi.org/10.1016/j.jecp.2005.01.004

Robinson, C. S., & Hayes, J. R. (1978). Making inferences about relevance in understanding problems. In R. Revlin & R. E. Mayer (Eds.), *Human reasoning* (pp. 195–206). New York, NY: Wiley.

St Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *The Quarterly Journal of Experimental Psychology, 59,* 745–759. http://dx.doi.org/10.1080/17470210500162854

Swanson, H. L. (2006). Cross-sectional and incremental changes in working memory and mathematical problem solving. *Journal of Educational Psychology, 98,* 265–281.

Swanson, H. L., & Sachse-Lee, C. (2001). Mathematical problem solving and working memory in children with learning disabilities: Both executive and phonological processes are important. *Journal of Experimental Child Psychology, 79,* 294–321. http://dx.doi.org/10.1006/jecp.2000.2587

Umemoto, T., Morikawa, Y., & Ibuki, M. (1955). The non-association values and meaningfulness of 1892 Japanese two-letter syllables and words. *The Japanese Journal of Psychology, 26,* 148–155.

Wu, L.-E. E. T. (1993). Japanese university entrance examination problems in mathematics. *American Educator: The Professional Journal of the American Federation of Teachers, 17,* 25–35.

Appendix A

Examples of Word Problems in Experiment 1

WP 1. There are five wooden pencils.
There are 2.6 times as many mechanical pencils as wooden pencils.
How many mechanical pencils are there?
WP 2. A rectangle is 28 cm long.
It is 26 cm wide.
What is its area?
WP 3. A rectangle is 28 cm long.
<u>A diagonal of this rectangle is 32 cm.</u>
This rectangle is 15 cm wide.
What is the rectangle's area?
WP 4. There are 2.5 l of soy sauce.
A mirin is 1.4 times as much soy sauce.
How many liters in the mirin?
WP 5. There are 23 apples.
There are 3 times as many oranges as the apples.
<u>There are 8 times as many grapes as apples.</u>
How many oranges are there?
WP 6. The distance to a destination is 18 km.
I move at 6 km per hour on foot.
How many hours do I take to get to the destination?
WP 7. There are four dogs.
<u>There are 5 more cats than dogs.</u>
There are 1.5 times as many birds as dogs.
How many birds are there?
WP 8. The distance to the destination is 30 km.
I move at 6 km per hour on foot.
<u>I move at 15 km per hour by bicycle.</u>
How long does it take me to get to the destination on foot?

Note. Underlined sentences are extraneous information. These examples were translated into English from Japanese.

(*Appendices continue*)

Appendix B

Examples of Word Problems in Experiment 2

WP 1. The distance to a destination is 48 km.
A snake crawls at 7 km/h.
How many hours does the snake take to get to the destination?

WP 2. A scallop weighs 120 g.
A cod roe weighs 20 g more than the scallop.
A soft seaweed weighs 30 g more than the scallop.
How much does the cod roe weigh?

WP 3. There are 4 hours to move.
A donkey does 19 km/h.
How far does the donkey move?

WP 4. A duck covers 2 km in an hour.
A gull covers 4 km in an hour.
They move for 6 h.
How far does the gull move?

WP 5. There are 9 oranges.
There are 3 times as many pears as oranges.
How many pears are there?

WP 6. A lily is 40 cm taller than a dandelion.
The dandelion is 30 cm.
How tall is the lily?

WP 7. A distance to a destination was 17 km.
A rabbit took 8 h to get to the destination.
A sheep took 6 h to get to the destination.
How many kilometers an hour did the rabbit cover?

WP 8. There are 3 times as many scissors as pencils.
There are 6 times as many seals as pencils.
There are 4 pencils.
How many seals are there?

Note. Underlined sentences are extraneous information. A word in necessary information was used as a necessary-information word in the lexical decision task. A word in extraneous information was used as an extraneous-information word. For example, *rabbit* was a necessary-information word and *sheep* was an extraneous-information word in WP 7.

Received December 15, 2014
Revision received April 14, 2016
Accepted April 14, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 2, 257–268

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000138

Effects of ABRACADABRA Literacy Instruction on Children With Autism Spectrum Disorder

Benjamin Bailey, Joanne Arciuli, and Roger J. Stancliffe
University of Sydney

This study explored the effects of ABRACADABRA, a free computer-assisted literacy program, on the reading accuracy and comprehension skills of children diagnosed with autism spectrum disorder (ASD). ABRACADABRA is a balanced literacy instruction program, targeting both code and meaning-based reading abilities. Twenty children with ASD, aged 5–11 years, were assigned by matched pairs to the instruction group or wait-list control group. Literacy instruction was delivered on a 1:1 basis in participants' homes over a 13-week period (26 sessions per participant). Pre and post instruction assessment using standardized measures revealed statistically significant gains in reading accuracy and comprehension for the instruction group relative to the wait-list control group, with large effect sizes. These findings indicate that children with ASD may benefit from ABRACADABRA literacy instruction.

Keywords: literacy, reading accuracy, reading comprehension, autism, ASD, ABRACADABRA

Early literacy skills provide a foundation for lifelong learning. Children who are skilled readers are more likely to experience positive academic outcomes and encounter fewer emotional and behavioral difficulties than their reading-delayed peers (Willcutt et al., 2007). These children also demonstrate greater motivation to complete academic tasks (Lyon, 1998) and are less inclined to leave school early, relative to less skilled readers (Daniel et al., 2006). In the longer term, skilled readers achieve more positive employment and economic outcomes (Roman, 2004), and exhibit greater health awareness than adults with poorer levels of reading ability (DeWalt, Berkman, Sheridan, Lohr, & Pignone, 2004). Given the potential benefits of skilled reading, there is an urgent need to establish effective literacy instruction for all children, including those with disabilities such as autism spectrum disorder (ASD).

ASD is an early onset developmental disability characterized by deficits in social communication, restricted patterns of interests, and engagement in repetitive behaviors (American Psychiatric Association, 2013). ASD is conceptualized as a spectrum disorder, meaning that these characteristics manifest heterogeneously throughout the population (Frazier et al., 2012). Global epidemiological research suggests that the median ASD prevalence rate is approximately 62/10,000 (Elsabbagh et al., 2012). ASD commonly co-occurs with difficulties in the areas of oral language (Lord, Risi, & Pickles, 2004), cognition (Chakrabarti & Fombonne, 2005), and behavior (Simonoff et al., 2008). Those more severely affected by ASD are more likely to present with associated comorbidities (Leyfer et al., 2006). The core characteristics of ASD, as well as the associated comorbid difficulties, can affect the literacy development of children within this population.

Reading and ASD

Reading is a dynamic process involving the interaction of two distinct components: decoding of text and comprehension of meaning (Gough & Tunmer, 1986). For both children without disabilities and children with ASD, these component reading abilities draw heavily on underlying cognitive and oral language skills (Jacobs & Richdale, 2013; Nation & Snowling, 2004). Given that ASD is often associated with deficits in cognition and oral language, it follows that some children with ASD are at increased risk of experiencing reading difficulties. Social-communicative deficits and behavioral difficulties may also restrict the ability of some children with ASD to adequately engage with literacy instruction, further impeding their reading development (Williams, Wright, Callaghan, & Coughlan, 2002).

In a seminal study of reading and autism, Nation and colleagues (2006) explored the reading accuracy and comprehension abilities of 41 children diagnosed with ASD (Nation, Clarke, Wright, & Williams, 2006). The researchers employed broad inclusion criteria, requiring only that participants were aged 6 to 15 years and had measureable oral language skills. Their analyses revealed that a considerable number of children exhibited difficulties in reading accuracy, with 22% of the participants completely unable to read single words and nonwords. Data from the remaining participants revealed an atypical profile of reading abilities characterized by relative strengths in reading accuracy and weaknesses in reading

This article was published Online First June 20, 2016.

Benjamin Bailey, Joanne Arciuli, and Roger J. Stancliffe, Centre for Disability Research and Policy, Faculty of Health Sciences, University of Sydney.

This work was partly supported by a Future Fellowship awarded to Joanne Arciuli by the Australian Research Council (FT130101570). Benjamin Bailey conducted this research as part of his PhD under the supervision of Joanne Arciuli (primary supervisor) and Roger J. Stancliffe (auxiliary supervisor). The authors thank the children and families who participated in this study. The authors also thank the staff at the Centre for the Study of Learning and Performance for their assistance in implementing the ABRACADABRA program in this study.

Correspondence concerning this article should be addressed to Joanne Arciuli, Faculty of Health Sciences, University of Sydney, PO Box 170, Lidcombe 1825, Australia. E-mail: joanne.arciuli@sydney.edu.au

comprehension. A comparable reading profile has been observed in recent studies involving children with ASD that have employed similar inclusion criteria (e.g., Arciuli, Stevens, Trembath, & Simpson, 2013).

A study by Nation and Snowling (1997) assessed reading difficulties in children without disabilities. Their data revealed strong correlations between the component reading abilities of reading accuracy and comprehension, and among subcomponent reading abilities, such as word level reading accuracy and passage level reading accuracy. Likewise, studies involving children with ASD have reported significant correlations between component and subcomponent reading abilities (Arciuli et al., 2013; Nation et al., 2006). However, the correlations found in these studies involving children with ASD tend to be lower than those reported in studies of children without disabilities. These findings indicate that the component and subcomponent reading abilities of children with ASD may develop more autonomously by comparison with children who do not have disabilities.

An emerging body of research addresses remediation of the reading difficulties exhibited by some children with ASD. In their review, Whalon, Al Otaiba, and Delano (2009) identified 11 studies, involving a total of 61 participants, which targeted some of the key reading-related abilities as defined by the National Reading Panel (NRP; National Institute of Child Health and Human Development [NICHD], 2000). Children with ASD were shown to benefit from instruction targeting phonics (e.g., Coleman-Martin, Heller, Cihak, & Irvine, 2005), oral reading fluency (e.g., Kamps, Barbetta, Leonard, & Delquadri, 1994), vocabulary (e.g., Kamps, Leonard, Potucek, & Garrison-Harrell, 1995), and instruction in comprehension strategies (e.g., Whalon & Hanline, 2008). However, few of the reviewed studies investigated whether these instructional approaches promoted the development of reading accuracy and comprehension skills in children with ASD. In addition, none of the reviewed studies evaluated the effects of comprehensive literacy instruction that targets all of the key reading-related abilities identified by the NRP. Given that children with ASD have been shown to benefit from some individual elements of literacy instruction, the evaluation of more comprehensive instructional approaches is of critical importance.

ABRACADABRA

ABRACADABRA (hereafter referred to as ABRA; Centre for the Study of Learning and Performance [CSLP], 2009) is a freely available literacy program designed to improve the reading and writing skills of all children, including those at risk of low literacy abilities. ABRA learning objectives are informed by the recommendations of the NRP (NICHD, 2000) and other reviews of effective reading interventions (see Abrami et al., 2010, for a description and explanation of the development of ABRA). Specifically, ABRA targets the development of foundational literacy skills including alphabetics, reading fluency, reading comprehension, and writing. Instruction targeting these skills is delivered using a combination of computer activities and noncomputerized extension tasks. According to Abrami et al. (2010), the pedagogical underpinnings of ABRA are intended to replicate those of balanced literacy programs, as described by Chall (1967) and Adams (1990). That is, ABRA learning activities emphasize a balance between children's code (i.e., phonics and word study) and meaning-based skill development (i.e., reading comprehension), and engagement with real literature.

ABRA is the focus of an ongoing research program at the CSLP at Concordia University. A recent meta-analysis identified nine randomized control trials and quasi-experimental studies that have examined the effects of ABRA on literacy outcomes as defined by the NRP (Abrami, Borohkovski, & Lysneko, 2015). These studies included Kindergarten, Grade 1, and Grade 2 children from diverse populations. There was no mention of children with disabilities. For example, research conducted by Wolgemuth et al. (2013) included indigenous Australian children, and the study conducted by Abrami et al. (2014) was undertaken in Sub-Saharan Africa. Across studies, students received between 10 to 32 hours of ABRA instruction in small groups or whole class settings for periods ranging from 8 to 16 weeks. Generally, these previous studies utilized standardized measures to assess outcomes. The results of the meta-analysis revealed that children who received ABRA instruction exhibited statistically significant gains in phonemic awareness, phonics, vocabulary, and listening comprehension compared with children in control conditions, with small effect sizes. Improvements in reading accuracy, reading comprehension, and reading fluency were evident in some previous studies; however, these gains did not always reach statistical significance. Divergent findings across some of the previous studies may be attributed to differences in the implementation of ABRA (e.g., small group vs. whole class administration of ABRA; differences in hours of instruction).

Computer-Assisted Instruction

Pedagogical approaches, such as ABRA, which utilize computer-assisted instruction (CAI) may be well suited to children with ASD (Grynszpan, Weiss, Perez-Diaz, & Gal, 2014). Unlike teacher-directed instruction, CAI is not heavily contingent upon social communicative abilities, which are a key deficit for children in this population (Williams et al., 2002). Previous research has shown that children with ASD tend to be more responsive during CAI that targets social, language, and communication development as compared with teacher directed approaches (Ploog, Scharf, Nelson, & Brooks, 2013).

A recent review evaluated the use of CAI for the teaching of reading and related skills for children with ASD (Ramdoss et al., 2011). Twelve studies were included in the review, involving a total of 94 participants. Evidence supported the use of CAI to develop skills associated with reading, including phonological awareness (e.g., Heimann, Nelson, Tjus, & Gillberg, 1995), receptive language (e.g., Whalen et al., 2010), vocabulary (e.g., Moore & Calvert, 2000), and sentence construction (e.g., Basil & Reyes, 2003), as well as component reading skills, decoding (e.g., Tjus, Heimann, & Nelson, 1998), and reading comprehension (e.g., Basil & Reyes, 2003). On average, analyses revealed large effects for these CAI programs. However, there was considerable variability across studies, with some showing CAI to be no more beneficial for children with ASD than teacher-led instruction (e.g., Travers et al., 2011).

Several issues need to be considered when evaluating the effects of CAI in children with ASD. Previous studies have often involved small samples, many comprised of less than 10 participants in total. Thus, some studies may have lacked the statistical power to

draw definitive conclusions. In addition, some previous studies have utilized CAI programs that are difficult to access (e.g., Coleman-Martin et al., 2005), costly (e.g., Whalen et al., 2010), and/or are now considered outdated (e.g., Heimann et al., 1995). Thus, the real-world applicability of many of the CAI programs that have been evaluated is questionable. Finally, many previous studies have relied on nonstandardized measures of literacy to evaluate outcomes following CAI (e.g., Whitcomb, Bass, & Luiselli, 2011). The lack of standardized measures in the previous research limits the generalizability of some of these studies.

The Current Study

In the current study we sought to explore the effects of ABRA on the reading skills of a diverse group of children with ASD. Literacy instruction was delivered using ABRA's freely available web application and noncomputerized extension tasks. Unlike previous ABRA research, the current study was conducted independently of the CSLP.

The study was guided by three research questions.

1. Can ABRA instruction improve the reading skills of children with ASD when compared with a control group of children with ASD who do not receive ABRA instruction?

2. Are improvements in reading ability following ABRA instruction observed across both word and passage levels for children with ASD?

3. How large are the improvements in reading ability following ABRA instruction for children with ASD?

We hypothesized that participants with ASD would exhibit improved reading accuracy and comprehension abilities following 13 weeks of ABRA instruction compared with a wait-list control group of children with ASD. In addition, we hypothesized that the relative gains achieved by participants with ASD following ABRA instruction would be observed across three aspects of reading ability: word level accuracy, passage level accuracy, and passage level comprehension. However, we were unsure about the size of these gains.

Method

Design

The study followed a pretest/posttest control group design. Participants were assigned to one of two experimental conditions: the wait-list control group or the instruction group. Pairs of participants who were of similar age and had comparable oral language, reading and adaptive abilities were identified. Participants in each pairing were then randomly assigned to opposing experimental conditions (i.e., the wait-list control or instruction group). These groupings were later altered slightly to accommodate changes in participant availability as advised by parents (i.e., one participant was removed from the instruction group and two participants were added in their place).

Participants in the instruction group received home-based 1:1 ABRA instruction over a period of 13 weeks (26 sessions per participant). Participants in the wait-list control group continued their normal academic schedule during this time. Thus, ABRA was supplemental for the instruction group while the wait-list control group went about their school activities "business as usual." Information was not collected regarding participants' normal school literacy instruction. Pre- and postinstruction assessment was carried out at the University or in the participant's home within 9 days of the instruction period. All assessment and instruction sessions were conducted by the first author who is a certified practicing speech pathologist with previous experience working with children on the autism spectrum.

Participants

Research advertisements were circulated throughout speech pathology and psychology clinics across a large metropolitan area within Sydney, Australia. The research protocol was approved by the relevant University's Human Research Ethics Committee. Legal guardians provided written informed consent prior to participation.

Eligibility for the study required that participants met the following inclusion criteria: (a) 5–11 years of age; (b) previous formal clinical diagnosis of ASD using Diagnostic and Statistical Manual (DSM) criteria; (c) no hearing or vision impairments; (d) measurable language ability; and (e) able to demonstrate sustained attention to tasks for 15 min. Of an initial pool of 25 participants, two were excluded because they did not meet the inclusion criteria. A further three participants were excluded because of conflicts in scheduling. Twenty children formed the final sample, of whom 18 were male. As expected, the final sample was highly heterogeneous and comprised of children with differing levels of developmental, adaptive and academic functioning. Participants were enrolled in inclusive education (i.e., classrooms with peers who do not have disabilities), support classes (i.e., classrooms with peers who have disabilities within a school for students without disabilities), or specialist settings (i.e., schools for children with ASD). Demographic and diagnostic information by group (wait-list control vs. instruction) is shown in Table 1.

Independent samples t tests with alpha set at .05 showed no statistically significant differences between the instruction and wait-list control groups for age, $t(18) = -3.54$, $p = .73$, and across baseline measures of adaptive ability, vocabulary, phonological

Table 1
Demographic and Diagnostic Information by Group

Characteristic	Wait-list control	ABRA instruction
Age[a]	90.22 (19.72)	87.18 (18.65)
Sex (M:F)	8:1	10:1
Reported diagnosis (ASD/Asp./PDD-NOS)	5:1:3	8:1:2
Secondary diagnoses (ADHD/LD/AD)	1:7:3	2:9:8
School (Inclusive/Support/Specialist)	6:3:0	8:1:2
>1 language spoken at home (Y/N)	3:6	5:6

Note. Asp. = Asperger's syndrome; PDD-NOS = pervasive developmental disorder-not otherwise specified; ADHD = attention-deficit/hyperactivity disorder; LD = language difficulties; AD = articulation difficulties; Inclusive = inclusive class; Support = support class; Specialist = specialist class. Data in parentheses are *SD*s.
[a] Age is reported in months.

awareness, word level reading accuracy, passage level reading accuracy, and passage level reading comprehension (see Table 2 for independent samples *t* test results). The percentile rank measures shown in Table 2 were calculated using normative data derived from samples that included a majority of children without disabilities (see the Measures section for further details). Normative data based solely on the ASD population were not available for any of the measures.

As expected, scores for most participants placed them well below the age-adjusted average on the measure of adaptive ability (i.e., only six participants achieved percentile rankings above the 16th percentile). Scores varied considerably within each group, and across each of the measures, reflecting the broad inclusion criteria utilized in the current study.

Measures

The measures used in the current study were selected for two purposes. First, tests of oral language, reading and adaptive ability were used to obtain baseline measures in order to assign participants to either the wait-list control or instruction group. Second, tests of word and passage level reading accuracy and comprehension were used to evaluate outcomes following ABRA instruction. All of the standardized tests included in the protocol are widely used, valid and reliable measures. Each provides age or year-of-schooling referenced percentile ranks. With the exception of adaptive ability, all assessments were administered individually to each participant by the first author. The measure of adaptive ability was obtained individually via semistructured parent interview with the first author. Participants received a score of zero if unable to satisfy basal level performance criteria on a test.

Where known, we report the percentage of children with ASD in the normative sample associated with each assessment. Some normative samples included children with ASD but considered these children as belonging to broader disability classifications. For these samples, the percentage of children in autism-related classifications is reported. For the remaining normative samples, the percentage of children with disabilities is reported.

Adaptive ability. Each parent participated in a semistructured interview using the Survey Interview Form from the Vineland Adaptive Behavior Scales–2nd Edition (VABS-2; Sparrow, Cicchetti, & Balla, 2005). The test evaluated the domains of communication (receptive, expressive, written), daily living skills (personal, domestic, community), and socialization (interpersonal relationships, play and leisure time, coping skills). Additional items measuring fine and gross motor skills were administered to parents of participants aged six years and younger (n = 10). Children with health impairments, traumatic brain injury, multiple impairments, and/or autism comprised 1.7% of the VABS-2 normative sample. Thus, most of the children included in the normative sample did not have disabilities. For the current sample, the VABS-2 was found to have a high level of internal consistency for children aged seven years and older (Cronbach's alpha = .97), and children aged six years and younger (Cronbach's alpha = .99).

Vocabulary. The Peabody Picture Vocabulary Test–4th edition (PPVT-4; Dunn & Dunn, 2007) is a test of receptive vocabulary. Using Form A, participants were instructed to select one of four images best illustrating a target word verbally presented by the researcher. The PPVT-4 includes many simple items to improve measurement of lower functioning and younger children. Children with ASD constituted 0.2% of the PPVT-4 normative sample. The vast majority of the remaining children did not have disabilities. For the current sample, the PPVT-4 was found to have a high level of internal consistency (Cronbach's alpha = .98).

Phonological awareness. The Phonological Awareness Composite Score (PACS) from the Comprehensive Test of Phonological Processing – 2nd Edition (CTOPP – 2; Wagner, Torgesen, Rashotte, & Pearson, 2013) was used to assess phonological awareness. The PACS is comprised of three related subtests: (a) Elision, which is a sound deletion task that measures ability to segment and manipulate sounds within words; (b) Blending Words, where participants listened to a series of audio-recorded sounds and were then required to blend these sounds together to form a whole word; and (c) Sound Matching (for participants aged five to six years only), in which participants identified one picture from a choice of three that began with the same sound as a word read by the researcher. Participants aged 7 to 11 years (n = 10) completed a Phoneme Isolation task, where they were instructed to identify the phoneme occupying a specified position in a target word. Children with learning or health impairments constituted approximately 5% of the CTOPP-2 normative sample—the remainder of the sample did not have disabilities. For the current sample, the CTOPP-2 was shown to have a high level of internal

Table 2
Mean Age-Based Percentile Rank for Each Preinstruction Measure by Group

	Wait-list control (n = 9)			ABRA instruction (n = 11)					
Measure	M	SD	Range	M	SD	Range	t(18)	p	Cohen's d
Adaptive ability	17.56	25.74	1–84	18.36	19.93	2–63	.08	.94	.04
Vocabulary	29.14	30.83	.3–79	26.00	19.45	1–53	.29	.78	.12
Phonological awareness	16.11	17.93	0–39	15.09	20.04	0–63	.12	.91	.05
Word level reading accuracy	38.89	35.44	2–87	43.27	31.61	2–98	.29	.77	.13
Passage level reading accuracy[a]	19.89	25.90	0–65	25.45	26.77	0–81	.47	.64	.21
Reading comprehension[a]	16.67	26.98	0–68	15.55	17.28	0–53	.11	.91	.05

Note. Adaptive ability: Vineland Adaptive Behavior Scales (VABS-2), Adaptive Behavior Composite; Vocabulary: Peabody Picture Vocabulary Test (PPVT-4); Phonological awareness: Comprehensive Test of Phonological Processing (CTOPP-2), Phonological Awareness Composite Score; Word level reading accuracy: Wide Range Achievement Test (WRAT-4), Word Identification subtest; Passage level reading accuracy and reading comprehension: Neale Analysis of Reading Ability (NARA-3).
[a] Data for passage level reading accuracy and reading comprehension are year-of-schooling based percentile ranks.

consistency for children aged six years and younger (Cronbach's alpha = .97), and children aged seven years and older (Cronbach's alpha = .98).

Word level reading accuracy. The Word Reading subtest of the Wide Range Achievement Test–4th Edition (WRAT-4; Wilkinson & Robertson, 2006) was used to measure participants' ability to accurately decode letters and words. Participants were directed to read aloud a list of individual letters followed by a list of real words. Word reading targets were arranged in order of least (e.g., "cat") to most difficult (e.g., "usurp"). Children with disabilities constituted approximately 5% of the WRAT-4 normative sample. The remaining children in the normative sample did not have disabilities. For the current sample, the WRAT-4 was found to have a high level of internal consistency (Cronbach's alpha = .95).

Passage level reading accuracy. The Reading Accuracy Composite Score from the Neale Analysis of Reading Ability–3rd edition (NARA-3; Neale, 1999) was used to assess participants' ability to accurately decode passage level text. This assessment required participants to read aloud a series of passages of increasing length and complexity. The NARA-3 manual does not report the number of children with ASD, or other disabilities, in its normative sample. For the current sample, the reading accuracy composite was found to have high internal consistency (Cronbach's alpha = .82).

Passage level reading comprehension. The Reading Comprehension Composite Score from the NARA-3 (Neale, 1999) was used to assess participants' ability to derive meaning from written text at the passage level. This involved participants reading a series of passages aloud before being asked a number of prescribed questions related to the text. For the current sample, the reading comprehension composite was shown to have high internal consistency (Cronbach's alpha = .95).

Procedure

Preinstruction assessment. Participants completed a standardized assessment battery of reading, oral language and adaptive abilities (see Measures section for a description of these assessments). The battery was necessary because we wanted to make sure that the groups were equivalent prior to one group receiving instruction. Assessment tasks were administered in the order in which they appear in the preceding section. Assessment sessions ranged from 60- to 90-min duration, depending on the abilities and behaviors of individual participants.

ABRACADABRA instruction. ABRA was implemented as per the standard recommended protocol with the exception of two purposeful adaptations, which were discussed with and approved by the CSLP. First, as a consequence of the 1:1 setting used in the current study, ABRA instruction sessions did not include collaborative work with child peers. Instead, additional time was assigned to the computer activities and a reward task at the end of the session, and participants worked collaboratively with the first author during the ABRA extension tasks (e.g., taking turns reading pages of a story). Second, in anticipation that some children with ASD would perform less consistently than children without disabilities, the criterion used to identify skill mastery was lowered slightly from 90% to 85% accuracy (further details regarding skill mastery are provided below).

ABRA activities targeted four key literacy abilities: (a) alphabetics, (b) reading fluency, (c) reading comprehension, and (d) writing (Table 3). Word level activities used to promote alphabetics (i.e., the ability to associate sounds with letters and use these sounds to create words) were presented in a hierarchical sequence. The sequence began with early developing skills (e.g., sound matching) and ended with more complex tasks (e.g., word segmentation and blending). Word attack skills targeted during the word level computer tasks were incorporated into passage level reading fluency and comprehension tasks. For example, participants could click on unfamiliar words in the passage level prediction task and observe them being decoded. Writing tasks required participants to type word and passage level targets on a computer to dictation. Within most activities, skill development and task autonomy were targeted using a system of least (e.g., encouraging independent decoding) to most (e.g., demonstrated decoding) prompts. Reward contingencies (e.g., shots in a hockey-themed comprehension game) were used to encourage ongoing participant motivation and engagement.

ABRA's balanced curriculum and graded learning tasks permitted highly individualized literacy instruction. The preinstruction assessment data was used to inform the researchers of each participant's profile of literacy abilities (i.e., relative strengths and weaknesses). These profiles were used in conjunction with the ABRA manual to identify learning objectives, tasks, and task difficulty settings appropriate for instruction. Learning objectives, tasks, and associated task difficulty settings were reviewed following each instruction session using ongoing measures of participant performance. A performance criterion of 65%–85% accuracy was employed to identify tasks of appropriate content and difficulty for instruction. Skill mastery was set at 85% accuracy for each independent task, maintained over three consecutive sessions.

Instruction consisted of two 60-min training sessions delivered weekly over a 13-week period working 1:1 with participants. Instruction sessions were conducted outside of school hours, and therefore necessarily in participants' homes, to minimize disruption to school activities. Computer activities were presented on a 15.6" laptop with participants seated one meter from the screen at eye level. These activities were designed to encourage independent participation (e.g., animated videos demonstrating task completion appeared prior to each activity). However, the experimenter was present for the duration of each session and assisted participants in transitioning between tasks. All participants had at least some ability to independently operate a standard computer mouse prior to commencing ABRA instruction. Some participants received additional support, in the form of hand-over-hand assistance, to operate the hardware during tasks which required rapid responses. Breaks were provided to participants as required throughout the instruction sessions.

Each 60-min ABRA session followed a routine structure. First, participants completed a 15-min computer task targeting word level abilities (i.e., alphabetics, high-frequency word identification, or word spelling). Next, participants completed a 20-min computer task targeting passage level abilities (i.e., reading fluency, reading comprehension, or sentence spelling). Skills targeted during the computer activities were then revisited during a 15-min, noncomputerized extension task which involved interaction between the experimenter and participant (e.g., shared reading or spelling games). Consistent with previous ABRA research, these extension tasks were guided by the

Table 3
ABRACADABRA Activities

Literacy domain	Task level	Task name	Task description
Alphabetics	All alphabetics tasks were word level	Matching sounds	Identify matching sounds
		Alphabet song	Sing along to the alphabet song
		Word counting	Count words in an audio-recording
		Syllable counting	Count syllables in an audio-recording
		Same word	Identify same vs. different words
		Same phoneme	Identify same vs. different phonemes
		Word matching	Identify words with same vs. different initial or final phoneme
		Animated alphabet	Watch animations featuring letter sounds, a letter-writing cue and an alliterative phrase for each letter of the alphabet
		Letter sound search	Identify letters corresponding to audio-recorded phonemes
		Letter identification bingo	Identify letters by name
		Rhyme matching	Identify pairs of rhyming words
		Word families	Substitute initial letter(s) to form a new word (e.g., map → mat → bat)
		Auditory blending	Match phonemically segmented word to image
		Auditory segmenting	Match audio-recording of full word to segmented version of target word
		Blending train	Identify target word following phonemically segmented audio-recording
		Basic decoding	Decode written word and match to corresponding image
		Word changing	Substitute letters to form a new word (e.g., rat → mat → map)
Reading fluency	Word level	High frequency words	Identify a list of high frequency words
	Passage level	Tracking	Scan passage level text from left to right
	Passage level	Expression	Identify audio-recording as being read with good vs. bad expression then read the same passage aloud with appropriate expression
	Passage level	Accuracy	Read passage of text without error
	Passage level	Speed	Read passage of text at appropriate pace
Reading comprehension	All reading comprehension tasks were passage level	Prediction	Predict future events during passage level narrative
		Comprehension monitoring	Identify words incorrectly substituted in the text
		Sequencing	Place story images in linear order following reading
		Summarizing	Respond to questions during passage level reading task (questions designed to highlight important plot elements)
		Vocabulary	Select sentences containing correct use of a target word
		Vocabulary (ESL)	Match audio-recorded words to corresponding images. Participants then included target words in a cloze passage.
		Story response	Respond verbally to questions following reading
		Story elements	Respond to multiple choice questions following reading
Writing	Word level	Spelling words	Words typed to dictation
	Passage level	Spelling sentences	Sentences typed to dictation

recommendations of the ABRA manual. At the end of each session, participants were rewarded with a 10-min free choice activity (e.g., Legos).

Postinstruction assessment. The postinstruction assessment included three outcome measures: (a) word level reading accuracy, (b) passage level reading accuracy, and (c) passage level reading comprehension.

Implementation Fidelity

Implementation fidelity was addressed across three levels: context, compliance, and competence fidelity (Fixsen, Naoom, Blase, Friedman, & Wallace, 2005). Context fidelity requires that the precursors necessary for effective instruction are in place prior to a program's implementation. In the current study, the first author ensured high context fidelity prior to beginning instruction by gaining access to the ABRA learning materials and the ABRA Learning Tool Kit Teacher Guide (hereafter referred to as the ABRA manual; Abrami, White, & Wade, 2010), and by completing ABRA administration training. ABRA administration training comprised two sessions conducted by a representative from the CSLP. During these training sessions, the first author and CSLP representative discussed the theoretical, developmental and pedagogical underpinnings of ABRA. The CSLP rep-

Table 4
Mean Raw Scores Pre- and Postinstruction for Each Outcome Measure by Group

Measure	Wait-list control (n = 9)			ABRA instruction (n = 11)		
	M	*SD*	Range	*M*	*SD*	Range
Preinstruction						
Word level reading accuracy	25.33	12.29	8–45	25.82	10.80	8–43
Passage level reading accuracy	20.11	23.60	0–64	20.09	18.64	0–45
Passage level reading comprehension	5.67	8.50	0–22	5.00	5.33	0–15
Postinstruction						
Word level reading accuracy	24.89	12.24	7–47	28.64	10.21	13–43
Passage level reading accuracy	19.44	22.75	0–65	25.82	21.29	1–58
Passage level reading comprehension	5.89	8.94	0–25	8.64	7.89	1–25

Note. Word level reading accuracy: Wide Range Achievement Test (WRAT-4), Word Identification subtest; Passage level reading accuracy and reading comprehension: Neale Analysis of Reading Ability (NARA-3).

resentative also presented information relevant to the administration of ABRA in the current study (e.g., the modifications necessary to accommodate 1:1 instruction) and demonstrated the use of the program. Before and between the training sessions, the first author was required to complete readings, including previous ABRA research, ABRA manual (Abrami, White, & Wade, 2010) and the teacher's zone on the CSLP website, and become proficient in the use of the ABRA software.

Compliance fidelity refers to the degree to which the core elements of a program are utilized during its implementation. Consistent with the recommendations of the CSLP (as per the ABRA manual and ABRA administration training), instruction sessions were individually planned to include computer and noncomputerized learning tasks targeting a balance of code and meaning-based learning objectives. Each participant's progression through these learning objectives was documented to ensure that all children completed the prescribed 26 hr of instruction and that learning objectives complied with the recommendations of the CSLP (e.g., performance criterion of 85% accuracy was used to identify skill mastery). In these ways, written documents composed during the instruction period (i.e., session plans and session notes) show that the core elements of ABRA instruction, including instructional content and duration, were implemented in the current study.

Competence fidelity is the level of skill with which the core elements of a program are delivered during its implementation. The standardized nature of the ABRA computer activities goes some way toward ensuring competence fidelity in the current study. For example, preprogrammed video models that are embedded within the ABRA computer program itself ensured that participants received an appropriate introduction to each computerized ABRA activity. However, external measures relating to the first author's implementation of ABRA were not collected. As such, competence fidelity cannot be independently verified in the current study.

Results

Raw scores for each of the outcome measures are provided in Table 4. As can be seen, children in the wait-list control group maintained or showed slight decreases in their raw scores over time. By contrast, children in the instruction group showed increases in their raw scores.

As the children within each group were of different ages and grades we converted raw scores to percentile ranks. The effects of ABRA instruction on participants' reading performance were evaluated using a series of 2 × 2 analyses of variance (ANOVAs; Time × Group) with $\alpha = .05$. The dependent variable used in these analyses was either age-based percentile rank (for the measure of word level reading accuracy percentile rank is calculated based on age in months) or year-of-schooling referenced percentile rank (for the measures of passage level reading accuracy and comprehension percentile ranks are calculated based on grade). ANOVAs conducted using participants' raw scores are also reported.

Word Level Reading Accuracy

A statistically significant interaction effect was observed for Time × Group on the word level reading accuracy measure, $F(1, 18) = 5.73$, $p < .05$, with a large effect size, $\eta_p^2 = .24$.[1] As shown in Figure 1, scores for participants in the instruction group increased from pre- to postinstruction assessment, suggesting improved word level reading ability. By contrast, scores for the wait-list control group decreased between these two time points.[2] Analysis of raw scores revealed a similar result, Time × Group interaction: $F(1, 18) = 12.50$, $p < .01$, $\eta_p^2 = .41$.

Passage Level Reading Accuracy

Analysis of the passage level reading accuracy data revealed a statistically significant Time × Group interaction, $F(1, 18) = 10.50$, $p < .01$, with a large effect size, $\eta_p^2 = .37$. Figure 2 shows an increase in mean percentile rank for the instruction group, suggesting an improvement in passage level reading accuracy, while there was relatively little change in the reading scores of the wait-list control group. Analysis of raw scores showed a

[1] η_p^2 of .01 is considered to be a small effect size, .06 a medium effect size, and .14 a large effect size (Richardson, 2011).

[2] Note that the wait-list control group achieved very similar raw scores on the WRAT-4 at pre- and postinstruction assessment (25.33 vs. 24.89). The slight decrease in percentile rank (shown in Figure 1) is likely because of the particular norming method used in the WRAT-4 (i.e., norms are based on age in months). That is, for the wait-list control group, participants' raw scores at postinstruction assessment corresponded to slightly lower percentile rankings because these participants were not making the kind of progress that would be expected with increasing age as was seen in the normative sample (largely comprised of individuals without disabilities).

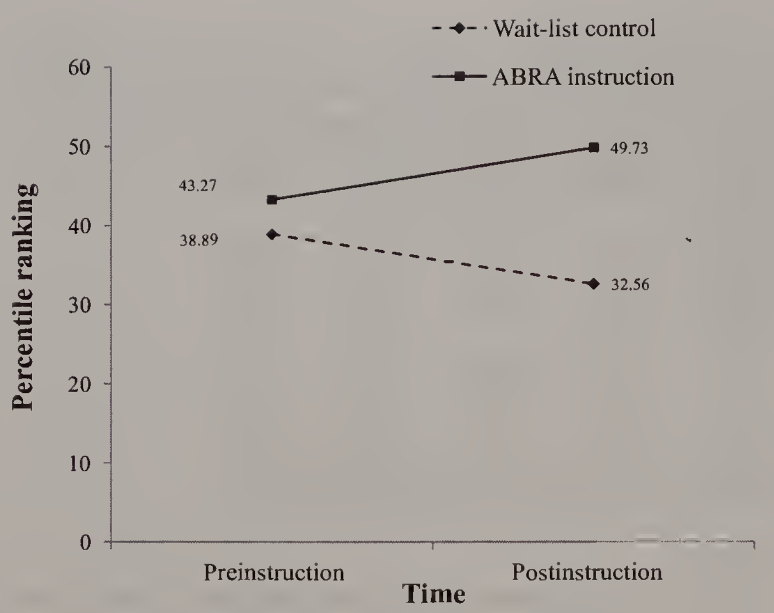


Figure 1. Mean percentile rankings for word level reading accuracy (Wide Range Achievement Test–4th Edition [WRAT-4]) by group.

similar result, Time × Group interaction: $F(1, 18) = 12.38$, $p < .01$, $\eta_p^2 = .41$.

Passage Level Reading Comprehension

A statistically significant Time × Group interaction was found for the measure of passage level reading comprehension, $F(1, 18) = 10.59$, $p < .01$, with a large effect size, $\eta_p^2 = .37$. As shown in Figure 3, percentile rank scores for the instruction group increased from pre- to postinstruction assessment, suggesting an improvement in passage level reading comprehension. Scores for the wait-list control group were relatively consistent across the two time points, indicating little change in reading comprehension skills. Analysis of participants' raw scores again revealed a similar result, Time × Group interaction: $F(1, 18) = 8.51$, $p < .01$, $\eta_p^2 = .32$.

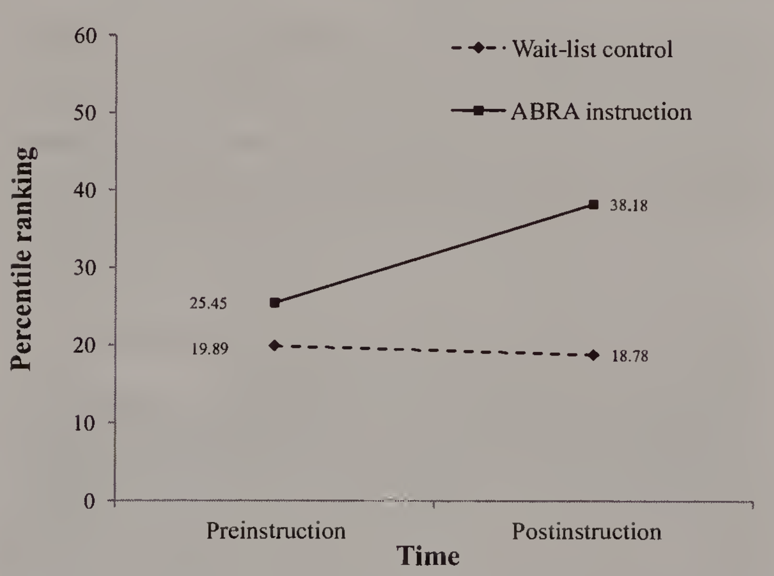


Figure 2. Mean percentile rankings for passage level reading accuracy (Neale Analysis of Reading Ability–3rd edition [NARA-3]) by group.

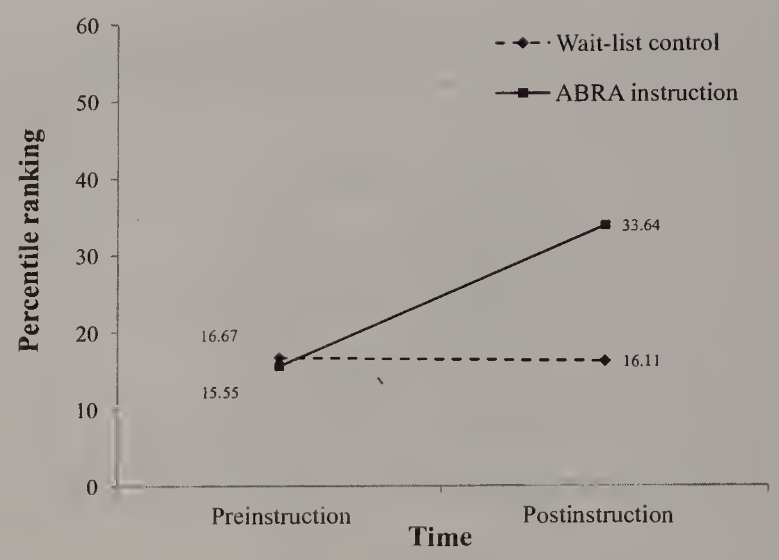


Figure 3. Mean percentile rankings for reading comprehension (Neale Analysis of Reading Ability–3rd edition [NARA-3]) by group.

Nonparametric Analyses

In view of the modest sample size, the effects of ABRA instruction were also evaluated using nonparametric Mann–Whitney tests conducted on pre-/postinstruction percentile rank difference scores for each of the outcome measures. The median difference score for the word level reading accuracy measure was 3 for the instruction group and −1 for the wait-list control group. For the passage level reading accuracy measure, the median difference score was 13 for the instruction group and 0 for the wait-list control group. For the passage level reading comprehension measure, the median difference score was 14 for the instruction group and 0 for the wait-list control group. With alpha set at .05, analyses showed statistically significant gains for the instruction group relative to the wait-list control group across all three reading measures: word level reading accuracy ($U = 21.00$, $z = -2.17$, $p = <.05$), passage level reading accuracy ($U = 10.50$, $z = -2.99$, $p = <.01$), and passage level reading comprehension ($U = 9.50$, $z = -3.07$, $p = <.01$).

Discussion

In the current study we examined the effects of ABRACADABRA literacy instruction on the reading abilities of a diverse group of children with ASD. Our research was guided by three questions intended to ascertain (a) whether ABRA instruction could be used to facilitate reading development in children with ASD; (b) whether the gains achieved using ABRA would be observed across both word and passage level reading abilities; and (c) the size of these gains.

We hypothesized that participants in the ABRA instruction group would exhibit improved reading abilities compared with a wait-list control group. Consistent with this hypothesis, participants in the instruction group, relative to the wait-list control group, achieved statistically significant gains in reading accuracy and comprehension following 26 sessions of ABRA instruction administered over a 13-week period. Our second hypothesis was that the relative gains achieved by participants in the instruction group would be observed across both word and passage level reading abilities. The data revealed statistically significant gains

for the instruction group, compared with the wait-list control group, across all three aspects of reading ability that were assessed (i.e., word level reading accuracy, passage level reading accuracy, and passage level reading comprehension), thus confirming our hypothesis.

With regard to our third hypothesis, effect size calculations showed considerable gains for the instruction group relative to the wait-list control group across each of the evaluated aspects of reading.[1] Gains achieved in the word level reading accuracy skills of the instruction group compared to the wait-list control group were large ($\eta_p^2 = .24$), suggesting that ABRA instruction was effective in facilitating substantial improvements in the word level reading abilities of children with ASD. By comparison with the wait-list control group, the instruction group also achieved large gains in their passage level reading accuracy ($\eta_p^2 = .37$) and passage level reading comprehension ($\eta_p^2 = .37$). It is interesting that gains in the instruction group were more pronounced with regard to participants' passage level reading skills as compared with their word level reading skills (as revealed by the ANOVAs that were conducted on percentile ranks—gains were equivalent in the ANOVAs that were conducted on raw scores). This uneven pattern of improvement may be further evidence that subcomponent reading skills can develop more autonomously in children with ASD as compared with children without disabilities (Arciuli et al., 2013; Nation et al., 2006).

Previous ABRA Research

The results presented here are in line with several studies showing that ABRA can have positive effects on children's reading abilities. Indeed, the effect sizes reported in the current study compare favorably with the previous ABRA research. For example, in contrast to the large effects reported in the current study, Wolgemuth et al. (2013) found ABRA instruction to have a statistically significant, medium-sized effect ($d = .36$) on the combined word reading accuracy and phonological awareness skills of children without disabilities. Other previous research has identified a modest average effect size ($g+ = 0.065$) for ABRA on the reading comprehension skills of children without disabilities (Abrami et al., 2015). Such comparisons suggest that children with ASD may be more receptive to ABRA instruction relative to children without disabilities. However, there are important differences in the way the ABRA literacy instruction was delivered in the current study versus previous studies. For instance, the current study evaluated the effects of ABRA instruction administered on a 1:1 basis whereas previous research has focused exclusively on the effects of ABRA instruction in small groups or whole class settings. Thus, comparison of effect sizes obtained from the current study and previous research should be carefully considered.

Numerous features within the ABRA program could potentially benefit children with ASD. Broadly speaking, it is posited that these features could contribute to the effectiveness of ABRA via children's improved engagement with instructional content, increased access to learning opportunities, and enhanced generalization of learned skills across instructional contexts.

Engagement. ABRA may serve to enhance the willingness and ability of children with ASD to engage with instructional content. ABRA sessions follow a set structure, which would appear well-suited to the needs of children with ASD in that they are commonly found to show a preference for repetition and predictability (Richler, Huerta, Bishop, & Lord, 2010). The interactive interface of ABRA's computer tasks is considered beneficial in that it requires children to actively engage with and respond to instruction. Active cognitive processing, such as that facilitated by ABRA, is critical to learning (Wouters, Paas, & van Merriënboer, 2008). ABRA activities also occur within the context of an overarching storyline. Embedding learning activities within a broader narrative in this way may enhance intrinsic motivation for learning (Baranowski, Buday, Thompson, & Baranowksi, 2008), and assist in the creation of an immersive learning environment (Dickey, 2006). Therefore the features of ABRA may help to reduce the difficulties some children with ASD have in engaging with instructional content.

Accessibility. ABRA instruction may promote the ability of children with ASD to access valuable literacy learning opportunities in several ways. First, learning objectives and task difficulty settings are tailored to ensure that each individual child commences instruction at an appropriate level and experiences the high rates of accurate responding necessary for efficient learning (Lamella & Tincani, 2012). Second, key reading skills and their associated learning tasks are introduced via animated video. This permits the use of visually cued instructions, the likes of which have been shown to benefit children with ASD (Quill, 1997). Third, many ABRA activities provide structured feedback using a system of least-to-most prompts. This form of feedback appears well-suited to children with ASD, many of whom prefer routine and may be averse to unpredictable feedback (Hume, Plavnick, & Odom, 2012). Considered collectively, these features are proposed to function in such a way as to assist children with ASD to access ABRA's instructional content despite their often considerable social-communicative, cognitive and behavioral difficulties.

Generalization. Our pre and post instruction testing utilized standardized assessments that were created independently of ABRA. Results revealed improvements for the instruction group relative to the wait-list control group. Thus, ABRA instruction appeared to generalize to a broader set of reading materials. ABRA's multimodal instructional approach may encourage the generalization of learnt reading skills for children with ASD in two ways. First, discrete reading skills, which are initially taught in isolation, are explicitly integrated into passage level reading tasks involving both decoding and reading comprehension. This form of embedded instruction may serve to enhance both the development of discrete skills and the abilities of children with ASD to independently apply these skills during novel tasks (Smith, Spooner, & Wood, 2013). Second, ABRA sessions are structured in such a way as to ensure that reading skills are targeted using both computer and noncomputerized learning tasks. The use of multiple mediums is proposed to aid in the development of generalized reading skills in children with ASD, many of whom are shown to have difficulty generalizing learned skills across instructional contexts (Hume, Loftin, & Lantz, 2009).

Previous CAI Research

The current study addressed some of the limitations in the previous research on CAI and ASD. These limitations include the use of small samples typically comprised of higher functioning children, reliance on nonstandardized outcome measures, and use

of CAI programs that are inaccessible, expensive, or outdated. We addressed these limitations by evaluating the effects of a freely accessible, computer-assisted program on the reading skills of a relatively large, diverse sample of children with ASD using standardized outcome measures. The inclusion of standardized outcome measures in the current study permitted us to directly compare participants of different ages and to quantify changes in reading ability for each participant with ASD from pre- to postinstruction with reference to a normative sample.

Previous research has returned mixed results regarding the effects of CAI on the reading skills of children with ASD. For example, Williams et al. (2002) found nonsignificant gains in word reading accuracy for children with ASD who received instruction using an experimental computer-based literacy program. By contrast, Tjus et al. (1998) reported significant improvements in the word and sentence reading accuracy skills of children with ASD following instruction using the Delta Messages program, with large effects (δ_{rm} = 1.031). Basil and Reyes (2003) also identified improvements in reading comprehension for a child with ASD following instruction using the Delta Messages program but did not report effect sizes.

It is important to note that the types of CAI investigated in previous research have differed widely in both instructional focus and mode of delivery. For example, where the current study utilized a balanced reading program (i.e., targeting both code and meaning based abilities) delivered using a web application and noncomputerized extension tasks, Tjus et al. (1998) administered a purely computerized intervention targeting only sentence construction skills. The instruction protocols employed across these studies have also differed in intensity and duration, ranging from a few days (e.g., Moore & Calvert, 2000) to several months (e.g., Bosseler & Massaro, 2003), and have involved divergent samples of children with ASD, differing widely in both age and level of functioning. Given these inconsistencies, it is difficult to directly compare the learning outcomes of children with ASD following exposure to the various CAI programs. However, the large effects reported in the current study suggest that ABRA may be among the more effective CAI programs for teaching reading skills to children with ASD.

Limitations and Future Research

While the findings reported in the current study are encouraging, several limitations warrant consideration. First, we did not collect information regarding the regular classroom literacy instruction that participants received during the instruction period. As a consequence, it is not possible to determine whether differences in classroom literacy instruction may have contributed to our results. However, given that participants in each group came from a number of different districts, we think it unlikely that classroom instruction could have had a systematic effect on the results. Second, assessment and instruction sessions were conducted by the first author. As such, it is possible that increased rapport may have affected the performance of participants in the instruction group at postinstruction assessment. However, we emphasize that pre and post instruction assessment utilized standardized measures with strict administration procedures, thereby limiting the effect of rapport. Third, external measures of competence fidelity were not collected. It is therefore unclear whether the first author implemented ABRA with a high degree of skill. However, there was strong evidence of context and compliance fidelity.

An evaluation of ABRA that addresses the above limitations with a larger sample of children with ASD is encouraged. A larger study would benefit from incorporating additional outcome measures (e.g., those relating to nonword decoding skills) and could explore the effects of ABRA on different subgroups within the ASD population, such as children with and without comorbid language difficulties. Future studies of children with ASD could also evaluate classroom-based or parent-directed administration of ABRA as well as the use of ABRA as a core, as opposed to supplemental, literacy program.

Conclusion

The current study is the first to evaluate the effects of ABRA instruction on children with ASD and, as far as we are aware, is the first investigation of ABRA to be conducted independently of the CSLP. Our findings demonstrate that children with ASD, like children without disabilities, can benefit from balanced literacy instruction that targets alphabetics, reading fluency, reading comprehension, and writing contained within the ABRA instruction program. The benefit we report here was observed across three aspects of reading ability: word level accuracy, passage level accuracy, and passage level reading comprehension. In short, the freely available, computer-assisted ABRA program shows great promise in improving reading outcomes for children with ASD.

References

Abrami, P. C., Borohkovski, E., & Lysneko, L. (2015). The effects of ABRACADABRA on reading outcomes: A meta-analysis of applied field research. *Journal of Interactive Learning Research, 26,* 337–367.

Abrami, P. C., Savage, R. S., Deleveaux, G., Wade, A., Meyer, E., & LeBel, C. (2010). The learning toolkit: The design, development, testing and dissemination of evidence-based educational software 1. In P. Zemliansky & D. Wilcox (Eds.), *Design and implementation of educational games: Theoretical and practical perspectives* (pp. 168–188). Hershey, PA: Information Science Reference. http://dx.doi.org/10.4018/978-1-61520-781-7.ch012

Abrami, P. C., Wade, A., Lysenko, L. V., Marsh, J., & Gioko, A. (2014). Using educational technology to develop early literacy skills in Sub-Saharan Africa. *Education and Information Technologies*. Advance online publication. http://dx.doi.org/10.1007/s10639-014-9362-4

Abrami, P. C., White, B., & Wade, A. (2010). *ABRACADABRA LTK teacher guide*. Retrieved from http://grover.concordia.ca/abracadabra/resources/download/LTK_ABRA_lr.pdf

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Arciuli, J., Stevens, K., Trembath, D., & Simpson, I. C. (2013). The relationship between parent report of adaptive behavior and direct assessment of reading ability in children with autism spectrum disorder. *Journal of Speech, Language, and Hearing Research, 56,* 1837–1844. http://dx.doi.org/10.1044/1092-4388(2013/12-0034)

Baranowski, T., Buday, R., Thompson, D. I., & Baranowski, J. (2008). Playing for real: Video games and stories for health-related behavior change. *American Journal of Preventive Medicine, 34,* 74–82. http://dx.doi.org/10.1016/j.amepre.2007.09.027

Basil, C., & Reyes, S. (2003). Acquisition of literacy skills by children with severe disability. *Child Language Teaching and Therapy, 19,* 27–48. http://dx.doi.org/10.1191/0265659003ct242oa

Bosseler, A., & Massaro, D. W. (2003). Development and evaluation of a computer-animated tutor for vocabulary and language learning in children with autism. *Journal of Autism and Developmental Disorders, 33,* 653–672. http://dx.doi.org/10.1023/B:JADD.0000006002.82367.4f

Centre for the Study of Learning and Performance. (2009, January). The learning toolkit (Version 2.27) [Web application]. Montreal, Canada: Concordia University. Retrieved from http://doe.concordia.ca/cslp/ICT-LTK.php

Chakrabarti, S., & Fombonne, E. (2005). Pervasive developmental disorders in preschool children: Confirmation of high prevalence. *The American Journal of Psychiatry, 162,* 1133–1141. http://dx.doi.org/10.1176/appi.ajp.162.6.1133

Chall, J. S. (1967). *Learning to read: The great debate.* New York, NY: McGraw-Hill.

Coleman-Martin, M. B., Heller, K. W., Cihak, D. F., & Irvine, K. L. (2005). Using computer-assisted instruction and the nonverbal reading approach to teach word identification. *Focus on Autism and Other Developmental Disabilities, 20,* 80–90. http://dx.doi.org/10.1177/10883576050200020401

Daniel, S. S., Walsh, A. K., Goldston, D. B., Arnold, E. M., Reboussin, B. A., & Wood, F. B. (2006). Suicidality, school dropout, and reading problems among adolescents. *Journal of Learning Disabilities, 39,* 507–514. http://dx.doi.org/10.1177/00222194060390060301

DeWalt, D. A., Berkman, N. D., Sheridan, S., Lohr, K. N., & Pignone, M. P. (2004). Literacy and health outcomes: A systematic review of the literature. *Journal of General Internal Medicine, 19,* 1228–1239. http://dx.doi.org/10.1111/j.1525-1497.2004.40153.x

Dickey, M. (2006). Game design narrative for learning: Appropriating adventure game design narrative devices and techniques for the design of interactive learning environments. *Educational Technology Research and Development, 54,* 245–263. http://dx.doi.org/10.1007/s11423-006-8806-y

Dunn, L. M., & Dunn, D. M. (2007). *The Peabody Picture Vocabulary Test* (4th ed.). Minneapolis, MN: Pearson Assessments.

Elsabbagh, M., Divan, G., Koh, Y. J., Kim, Y. S., Kauchali, S., Marcín, C., . . . Fombonne, E. (2012). Global prevalence of autism and other pervasive developmental disorders. *Autism Research, 5,* 160–179. http://dx.doi.org/10.1002/aur.239

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature* [Monograph]. Retrieved from http://nirn.fpg.unc.edu/sites/nirn.fpg.unc.edu/files/resources/NIRN-MonographFull-01-2005.pdf

Frazier, T. W., Youngstrom, E. A., Speer, L., Embacher, R., Law, P., Constantino, J., . . . Eng, C. (2012). Validation of proposed *DSM–5* criteria for autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 51,* 28–40.e3. http://dx.doi.org/10.1016/j.jaac.2011.09.021

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *RASE: Remedial and Special Education, 7,* 6–10. http://dx.doi.org/10.1177/074193258600700104

Grynszpan, O., Weiss, P. L. T., Perez-Diaz, F., & Gal, E. (2014). Innovative technology-based interventions for autism spectrum disorders: A meta-analysis. *Autism, 18,* 346–361. http://dx.doi.org/10.1177/1362361313476767

Heimann, M., Nelson, K. E., Tjus, T., & Gillberg, C. (1995). Increasing reading and communication skills in children with autism through an interactive multimedia computer program. *Journal of Autism and Developmental Disorders, 25,* 459–480. http://dx.doi.org/10.1007/BF02178294

Hume, K., Loftin, R., & Lantz, J. (2009). Increasing independence in autism spectrum disorders: A review of three focused interventions. *Journal of Autism and Developmental Disorders, 39,* 1329–1338. http://dx.doi.org/10.1007/s10803-009-0751-2

Hume, K., Plavnick, J., & Odom, S. L. (2012). Promoting task accuracy and independence in students with autism across educational setting through the use of individual work systems. *Journal of Autism and Developmental Disorders, 42,* 2084–2099. http://dx.doi.org/10.1007/s10803-012-1457-4

Jacobs, D. W., & Richdale, A. L. (2013). Predicting literacy in children with a high-functioning autism spectrum disorder. *Research in Developmental Disabilities, 34,* 2379–2390. http://dx.doi.org/10.1016/j.ridd.2013.04.007

Kamps, D. M., Barbetta, P. M., Leonard, B. R., & Delquadri, J. (1994). Classwide peer tutoring: An integration strategy to improve reading skills and promote peer interactions among students with autism and general education peers. *Journal of Applied Behavior Analysis, 27,* 49–61. http://dx.doi.org/10.1901/jaba.1994.27-49

Kamps, D. M., Leonard, B., Potucek, J., & Garrison-Harrell, L. (1995). Cooperative learning groups in reading: An integration strategy for students with autism and general classroom peers. *Behavioral Disorders, 21,* 89–109. Retrieved from http://www.jstor.org/stable/23888333

Lamella, L., & Tincani, M. (2012). Brief wait time to increase response opportunity and correct responding of children with autism spectrum disorder who display challenging behavior. *Journal of Developmental and Physical Disabilities, 24,* 559–573. http://dx.doi.org/10.1007/s10882-012-9289-x

Leyfer, O. T., Folstein, S. E., Bacalman, S., Davis, N. O., Dinh, E., Morgan, J., . . . Lainhart, J. E. (2006). Comorbid psychiatric disorders in children with autism: Interview development and rates of disorders. *Journal of Autism and Developmental Disorders, 36,* 849–861. http://dx.doi.org/10.1007/s10803-006-0123-0

Lord, C., Risi, S., & Pickles, A. (2004). Trajectory of language development in Autistic Spectrum Disorders. In M. L. Rice & S. F. Warren (Eds.), *Developmental language disorders: From phenotypes to etiologies* (pp. 7–29). Mahwah, NJ: Erlbaum Publishers.

Lyon, G. (1998). *Overview of reading and literacy initiatives.* Statement delivered to Committee on Labor and Human Resources, Washington, DC. Retrieved from http://files.eric.ed.gov/fulltext/ED444128.pdf

Moore, M., & Calvert, S. (2000). Brief report: Vocabulary acquisition for children with autism: Teacher or computer instruction. *Journal of Autism and Developmental Disorders, 30,* 359–362. http://dx.doi.org/10.1023/A:1005535602064

Nation, K., Clarke, P., Wright, B., & Williams, C. (2006). Patterns of reading ability in children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 36,* 911–919. http://dx.doi.org/10.1007/s10803-006-0130-1

Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology, 67,* 359–370. http://dx.doi.org/10.1111/j.2044-8279.1997.tb01250.x

Nation, K., & Snowling, M. J. (2004). Beyond phonological skills: Broader language skills contribute to the development of reading. *Journal of Research in Reading, 27,* 342–356. http://dx.doi.org/10.1111/j.1467-9817.2004.00238.x

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel: Teaching children to read.* Washington, DC: U.S. Government Printing Office.

Neale, M. D. (1999). *Neale Analysis of Reading Ability* (3rd ed., Australian Standardisation). Camberwell, Victoria, Australia: ACER Press.

Newschaffer, C. J., Croen, L. A., Daniels, J., Giarelli, E., Grether, J. K., Levy, S. E., . . . Windham, G. C. (2007). The epidemiology of autism spectrum disorders. *Annual Review of Public Health, 28,* 235–258. http://dx.doi.org/10.1146/annurev.publhealth.28.021406.144007

Ploog, B. O., Scharf, A., Nelson, D., & Brooks, P. J. (2013). Use of computer-assisted technologies (CAT) to enhance social, communica-

tive, and language development in children with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 43,* 301–322. http://dx.doi.org/10.1007/s10803-012-1571-3

Quill, K. A. (1997). Instructional considerations for young children with autism: The rationale for visually cued instruction. *Journal of Autism and Developmental Disorders, 27,* 697–714. http://dx.doi.org/10.1023/A:1025806900162

Ramdoss, S., Mulloy, A., Lang, R., O'Reilly, M., Sigafoos, J., Lancioni, G., . . . El Zein, F. (2011). Use of computer-based interventions to improve literacy skills in students with autism spectrum disorders: A systematic review. *Research in Autism Spectrum Disorders, 5,* 1306–1318. http://dx.doi.org/10.1016/j.rasd.2011.03.004

Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6,* 135–147. http://dx.doi.org/10.1016/j.edurev.2010.12.001

Richler, J., Huerta, M., Bishop, S. L., & Lord, C. (2010). Developmental trajectories of restricted and repetitive behaviors and interests in children with autism spectrum disorders. *Development and Psychopathology, 22,* 55–69. http://dx.doi.org/10.1017/S0954579409990265

Roman, S. P. (2004). Illiteracy and older adults: Individual and societal implications. *Educational Gerontology, 30,* 79–93. http://dx.doi.org/10.1080/03601270490266257

Simonoff, E., Pickles, A., Charman, T., Chandler, S., Loucas, T., & Baird, G. (2008). Psychiatric disorders in children with autism spectrum disorders: Prevalence, comorbidity, and associated factors in a population-derived sample. *Journal of the American Academy of Child & Adolescent Psychiatry, 47,* 921–929. http://dx.doi.org/10.1097/CHI.0b013e318179964f

Smith, B. R., Spooner, F., & Wood, C. L. (2013). Using embedded computer-assisted explicit instruction to teach science to students with autism spectrum disorder. *Research in Autism Spectrum Disorders, 7,* 433–443. http://dx.doi.org/10.1016/j.rasd.2012.10.010

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland Adaptive Behavior Scales* (2nd ed.). Circle Pines, MN: AGS Publishing.

Tjus, T., Heimann, M., & Nelson, K. E. (1998). Gains in literacy through the use of a specially developed multimedia computer strategy: Positive findings from 13 children with autism. *Autism, 2,* 139–156. http://dx.doi.org/10.1177/1362361398022003

Travers, J. C., Higgins, K., Pierce, T., Boone, R., Miller, S., & Tandy, R. (2011). Emergent literacy skills of preschool students with autism: A comparison of teacher-led and computer-assisted instruction. *Education and Training in Autism and Developmental Disabilities, 46,* 326–338. Retrieved from http://www.jstor.org/stable/23880589

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2013). *Comprehensive Test of Phonological Processing* (2nd ed.). Austin, TX: PRO-ED.

Whalen, C., Moss, D., Ilan, A. B., Vaupel, M., Fielding, P., Macdonald, K., . . . Symon, J. (2010). Efficacy of TeachTown: Basics computer-assisted intervention for the intensive comprehensive autism program in Los Angeles unified school district. *Autism, 14,* 179–197. http://dx.doi.org/10.1177/1362361310363282

Whalon, K. J., Al Otaiba, S., & Delano, M. E. (2009). Evidence-based reading instruction for individuals with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities, 24,* 3–16. http://dx.doi.org/10.1177/1088357608328515

Whalon, K. J., & Hanline, M. F. (2008). Effects of a reciprocal questioning intervention on the question generation and responding of children with autism spectrum disorder. *Education and Training in Developmental Disabilities, 43,* 367–387. Retrieved from http://www.jstor.org/stable/23879798

Whitcomb, S., Bass, J., & Luiselli, J. (2011). Effects of a computer-based early reading program (Headsprout) on word list and text reading skills in a student with autism. *Journal of Developmental and Physical Disabilities, 23,* 491–499. http://dx.doi.org/10.1007/s10882-011-9240-6

Wilkinson, G. S., & Robertson, G. J. (2006). *Wide Range Achievement Test* (4th ed.). Lutz, FL: Psychological Assessment Resources.

Willcutt, E. G., Betjemann, R. S., Pennington, B. F., Olson, R. K., DeFries, J. C., & Wadsworth, S. J. (2007). Longitudinal study of reading disability and attention-deficit/hyperactivity disorder: Implications for education. *Mind, Brain, and Education, 1,* 181–192. http://dx.doi.org/10.1111/j.1751-228X.2007.00019.x

Williams, C., Wright, B., Callaghan, G., & Coughlan, B. (2002). Do children with autism learn to read more readily by computer assisted instruction or traditional book methods? A pilot study. *Autism, 6,* 71–91. http://dx.doi.org/10.1177/1362361302006001006

Wolgemuth, J. R., Savage, R., Helmer, J., Harper, H., Lea, T., Abrami, P., . . . Louden, W. (2013). ABRACADABRA aids indigenous and non-indigenous early literacy in Australia: Evidence from a multisite randomized controlled trial. *Computers & Education, 67,* 250–264. http://dx.doi.org/10.1016/j.compedu.2013.04.002

Wouters, P., Paas, F., & van Merriënboer, J. J. (2008). How to optimize learning from animated models: A review of guidelines based on cognitive load. *Review of Educational Research, 78,* 645–675. http://dx.doi.org/10.3102/0034654308320320

Received July 7, 2015
Revision received April 8, 2016
Accepted May 2, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 2, 269–299

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000133

Creativity and Academic Achievement: A Meta-Analysis

Aleksandra Gajda and Maciej Karwowski
The Maria Grzegorzewska University

Ronald A. Beghetto
University of Connecticut

This article reports on a meta-analysis of 120 studies (total N = 52,578; 782 effects) examining the relationship between creativity and academic achievement in research conducted since the 1960s. Average correlation between creativity and academic achievement was r = .22, 95% CI [.19, .24]. An analysis of moderators revealed that this relationship was constant across time but stronger when creativity was measured using creativity tests compared to self-report measures and when academic achievement was measured using standardized tests rather than grade point average. Moreover, verbal tests of creativity yielded significantly stronger relationships with academic achievement than figural tests. Theoretical and practical consequences are discussed.

Keywords: creativity, school grades, academic achievement, meta-analysis

Supplemental materials: http://dx.doi.org/10.1037/edu0000133.supp

Is there a relationship between creativity and academic achievement? This is a longstanding and largely unresolved question. For more than half a century, educators and psychologists have attempted to address this issue (Cline, Richards, & Abe, 1962; Mednick, 1963). At a conceptual level, scholars have asserted that creativity and learning represent interrelated phenomena (e.g., Beghetto, 2016a; Guilford, 1967; Piaget, 1962, 1981; Sawyer, 2012; Vygotsky, 1967/2004). Some of the earliest and most prominent theorists in the field have noted this link. Guilford (1967), for instance, asserted that creativity and learning are essentially the same phenomenon. Vygotsky (1967/2004) similarly argued that the creative imagination "is a completely essential condition for almost all human mental activity" (p. 17). Another example is Piaget's theory of genetic epistemology. Indeed, creativity is central to Piaget's theory of learning. As Gruber (in Bringuier, 1980) has explained in reference to Piaget's theory, "The child does not learn simply what the adult tells him, he reinvents. It's a kind of creativity" (p. 67).

Regardless of the theoretical stance one takes on learning—be it behavioral, cognitive, constructivist, situated, sociomaterial, or some other theoretical orientation—creativity and learning share fundamental similarities. Indeed, both creativity and learning involve change. More specifically, creativity refers to new and meaningful changes in thoughts, products, and actions (Beghetto, 2016a; Sternberg, 1999). Similarly, learning represents relatively stable changes in understanding and behavior (Alexander, Schallert, & Reynolds, 2009). Moreover, both learning and creativity can be viewed as processes and products (e.g., Alexander et al., 2009; Beghetto, 2016a; Donovan & Bransford, 2005; Mumford, Medeiros, & Partlow, 2012; Wallas, 1926). It is therefore possible to say that "a creative act [as a product] is an instance of learning [as a process], for it represents a change in behavior" (Guilford, 1950, p. 446). Along these same lines, it is also possible to say learning (as a product) is a creative process, because it results from new and personally meaningful changes in one's prior understanding (Beghetto, 2016a). Given the theoretical links between creativity and learning, it seems reasonable to assume that there would be a positive relationship between creativity and measures of academic achievement. The empirical work that has examined this link, however, has yielded a more equivocal picture. Some researchers have, for instance, reported positive associations ranging from .10–.56 (Cicirelli, 1967; Getzels & Jackson, 1962; Niaz, Núñez, & Pineda, 2000; Ohnmacht, 1966). Others have reported little or no association (e.g., Edwards & Tyler, 1965; Grigorenko et al., 2009). Still others have reported negative associations (e.g., Anderson, White, & Stevens, 1969). In fact, some researches have noted all three patterns within the same study (e.g., Gralewski & Karwowski, 2012). Consequently, the best that can be said about whether there is a link between creativity and academic achievement is this: *It depends.*

Why might this be the case? The present meta-analysis endeavors to address this question. More specifically, we have two primary aims for our study. Our first goal is to provide an average effect size of the relationship between creativity and academic achievement. Our second goal is to examine the potential impact of factors that may moderate the relationship between creativity and academic achievement. Although there are examples of meta-analytic studies that have addressed related issues (e.g., the relationship between creativity and intelligence; see Kim, 2005), we are not aware of any published meta-analytic studies of creativity

This article was published Online First August 18, 2016.

Aleksandra Gajda and Maciej Karwowski, Department of Educational Sciences, The Maria Grzegorzewska University; Ronald A. Beghetto, Department of Educational Psychology, University of Connecticut.

The preparation of this article was supported by a grant from the National Science Centre, Poland (UMO-2011/03/N/HS6/05073).

Correspondence concerning this article should be addressed to Maciej Karwowski, Creative Education Lab, The Maria Grzegorzewska University, Szczesliwicka 40 Street, 02353 Warsaw, Poland. E-mail: mkarwowski@aps.edu.pl

and academic achievement.[1] We therefore endeavor to shed light on the mixed findings of prior research by providing a more stable estimate of the relationship between creativity and academic achievement and by examining factors that may potentially moderate this relationship.

Creativity and Academic Achievement

Creativity

Creativity scholars generally agree that creativity represents a combination between originality, novelty, or newness *and* usefulness, meeting task constraints, or meaningfulness as defined within a particular sociocultural and historical context (Amabile, 1996; Kaufman & Beghetto, 2009; Plucker, Beghetto, & Dow, 2004; Simonton, 2012; Sternberg & Lubart, 1999). The following simplified notation captures this definition (adapted from Beghetto & Kaufman, 2014; Simonton, 2012):

$$C = O \times TC$$

[—-CONTEXT—-]

In the above notation, C refers to creativity, O refers to originality, and TC refers to task constraints. As specified by this formulation, creativity is a multiplicative combination of originality and task constraints as situated within a particular context. Consequently, something that is original ($O = 1$) but does not meet contextually defined task constraints ($TC = 0$) could be called original but not creative ($C = 0$). Consider, for instance, a student taking a calculus exam who produces a vivid and quite stunning pencil drawing of mathematical symbols transforming into doves (instead of solving the problem presented on the exam). Such a response is clearly original, but it would not be considered creative in the context of the exam. In order for a student's response on a calculus exam to be considered creative, it would need to represent a novel solution to the problem at hand (i.e., meet the task constraints).

In the context of academic learning, creativity can be thought of as occurring at both a subjective (creativity as part of the act of learning) and an intersubjective (learning as a creative act) level (Beghetto, 2016a). At the subjective level, students exercise their creativity by developing new and personally meaningful ideas, insights, and understandings within the context of particular academic constraints (Beghetto, 2007; Beghetto & Kaufman, 2007). At the intersubjective level, students who share their unique and academically accurate insights and interpretations can also contribute to the learning and understanding of others (Beghetto, 2016a).

In this way, creativity is more than originality (Beghetto, 2010), divergent thinking (Baer, 1993; Beghetto, 2013; Guilford, 1967; Runco, 1991), or vividness of imagination (Dziedziewicz & Karwowski, 2015; Jankowska & Karwowski, 2015). It also involves deductive and inductive thinking (Dunbar, 1997; Vartanian, Martindale, & Kwiatkowski, 2003; Weisberg, 2006), as well as the ability to use specific problem-solving strategies to generate novel solutions to complex and ill-defined problems (Beghetto, 2016b; Finke, Ward, & Smith, 1992; Sternberg, 1998). All these characteristics are important for the acquisition of new knowledge and learning (Greiff et al., 2013). In this way, creativity and learning work hand-in-hand (e.g., Beghetto, 2016a; Guilford, 1967; Piaget, 1981; Vygotsky, 1967/2004). It therefore seems reasonable to suggest that creativity would be related to academic achievement, which is conceptualized as the outcome of learning.

Academic Achievement

Academic achievement is an outcome of learning, which is typically measured by classroom grades, classroom assessments, and external achievement tests. Researchers who have examined correlates of academic achievement have identified a wide array of factors, including individual, social, and sociocultural influences (see Hattie, 2009, for a review). Of these, student characteristics play one of the broadest and most influential roles in explaining variations in academic achievement. Student characteristics represent a highly heterogeneous dimension, which includes personality (Chamorro-Premuzic & Furnham, 2003; Poropat, 2009), cognitive abilities (e.g., Chamorro-Premuzic & Furnham, 2008; Deary, Strand, Smith, & Fernandes, 2007), intensity and type of motivation (Di Domenico & Fournier, 2015), self-esteem and academic self-concept (Marsh & Hau, 2004), and socioeconomic factors (Johnson, McGue, & Iacono, 2007; Sackett, Kuncel, Arneson, Cooper, & Waters, 2009).

Creativity is yet another student characteristic that shares a conceptual, albeit equivocal, link with academic achievement. As we have discussed, researchers have reported associations that are relatively strong (e.g., $r = .41$, Marjoribanks, 1976 or $r = .66$, Yeh, 2004), modest (e.g., $r = .20$, McCabe, 1991), null (e.g., $r = .03$, Tatlah, Aslam, Ali, & Iqbal, 2012), and, in some cases, negative (e.g., $r = -.03$, Anderson et al., 1969). The aim of the present study is to help clarify the empirical ambiguity surrounding the link between creativity and academic achievement by providing a stable estimate of the association and also examine whether and how potential moderators might influence that association.

Potential Moderators

What might account for variations in the relationship between creativity and academic achievement? Researchers who have addressed this question (e.g., Freund & Holling, 2008; Gralewski & Karwowski, 2012; Vijetha & Jangaiah, 2010) have identified several moderating factors (see Figure 1). As illustrated in Figure 1, those factors include (a) the type of measurement used, (b) grade level of participants, (c) the decade the study was conducted, and (d) the geographic region of the study. In the sections that follow, we briefly describe each of these potential moderators.

Type of Measurement

The type of measurement represents one of the most clearly identifiable moderators of the empirical relationship between cre-

[1] An anonymous reviewer brought to our attention an unpublished report (Halliburton-Beatty & Simms, 2013) that reanalyzed meta-analytic data testing the impact of creativity training programs on school achievement (presented in the meta-synthesis by Hattie, 2009). This report, however, has different scope and focus than our present study. As previously described, our analysis focuses on the relationship between creative ability/self-concepts and academic achievement (rather than the impact of creative training programs), and we analyze effects reported in primary source material (rather than a reanalysis).

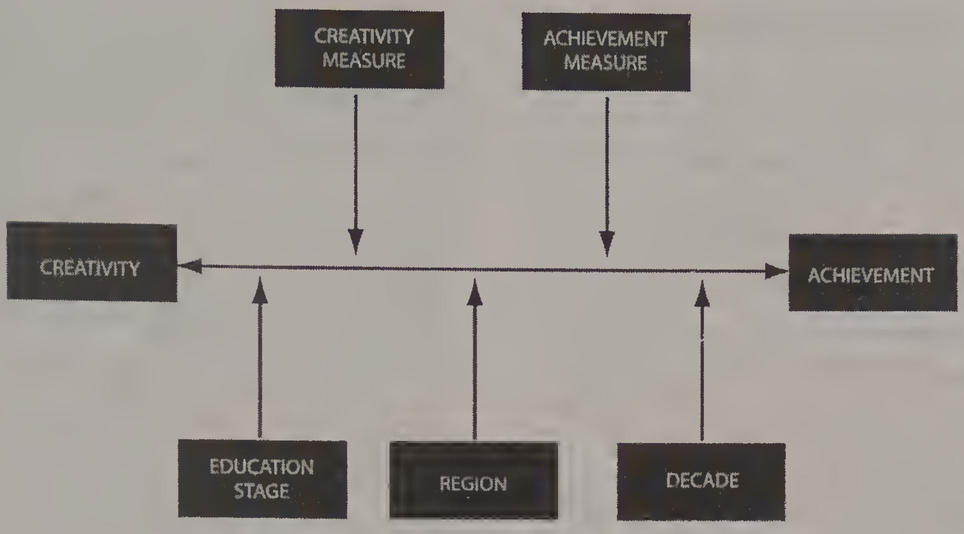


Figure 1. Potential factors influencing creativity and achievement relationship.

ativity and academic achievement. Put simply, regardless of the conceptual overlap between creativity and academic achievement, the degree of the observed relationship between creativity and academic achievement will, in large part, be determined by the amount of overlap in how each construct is measured. Moreover, there is a wide array of methods measures that can and have been used to measure both constructs (Freund & Holling, 2008; Gralewski & Karwowski, 2012). With respect to creativity, this includes everything from self-report measures to more objective creativity tests. To further compound this issue, there is little consensus in the field of how to best measure creativity (see Freund & Holling, 2008; Kaufman, Plucker, & Baer, 2008).

The kinds of creativity measures typically used to examine the relationship between creativity and achievement can be classified into two types: self-report and more objective creativity tests. Self-report measures tend to focus on beliefs about one's creative ability (e.g., Karwowski & Lebuda, 2015; Skager, Klein, & Schultz, 1967), creative activity or achievement (mainly inventories measuring the intensity of declared creative behaviors and activities or observable creative accomplishments, e.g., Carson, Peterson, & Higgins, 2005; Jauk, Benedek, & Neubauer, 2014), and indicators of creative personality (e.g., Naderi, Abdullah, Aizan, Sharir, & Kumar, 2009). More objective creativity tests tend to focus on divergent thinking skills (i.e., the ability to generate original ideas). These include tests based on Guilford's theory (e.g., Toll, 1985), the Test of Creative Thinking–Drawing Production (TCT-DP) by Urban and Jellen (Urban, 1991; Karwowski & Gralewski, 2013), the Torrance Test of Creative Thinking (TTCT; Clapham, 2004; Torrance, 1968), and other instruments (e.g., the Remote Associates Test, Mednick, 1963, or the Sternberg Triarchic Abilities Test, Chooi, Long, & Thompson, 2014).

Creativity tests can be further distinguished by modality: verbal tests (i.e., requiring participants to provide verbal answers to the problems provided; e.g., the TTCT verbal, Hansenne & Legrand, 2012, or the Verbaler Kreativity-Test, Rindermann & Neubauer, 2004) and figural tests (i.e., requiring participants to draw the solution; e.g., the TCT-DP, Gralewski & Karwowski, 2012, or the Test of Creative Imagery Ability, Jankowska & Karwowski, 2015). The most popular divergent thinking tests (e.g., TTCT) can be further divided into dimensions of divergent thinking (i.e., fluency, flexibility, originality, or elaboration). Previous studies have demonstrated that aspects of divergent thinking vary in their association with academic achievement (e.g., Auzmendi, Villa, & Abedi, 1996; Feldhusen, Treffinger, Van Mondfrans, & Ferris, 1971). We therefore explore whether these different dimensions influence the relationship between creativity and academic achievement but, given the limited work in this area, have no prediction as to the specific strength of this influence (e.g., nonexistent, weak, moderate, strong).

With respect to academic achievement, researchers have also used a wide array of methods and measures to examine the relationship with creativity. Similar to creativity measures, academic achievement measures can be classified into two types: subjective assessments and objective tests. Grade point averages (GPAs) represent the most common type of subjective measure used in studies that have examined the link with creativity (e.g., Chamorro-Premuzic, 2006; Freund & Holling, 2008; Gralewski & Karwowski, 2012). More objective tests refer to any externally constructed tests of academic subject matter knowledge or achievement (e.g., Tan, Mourgues, Bolden, & Grigorenko, 2014).

Taken together, the measures of creativity and academic achievement typically used in studies that have examined their relationship tend to include both subjective and more objective types of measurement. Moreover, creativity measures tend to focus more on assessing divergent thinking skills and abilities (e.g., generating original ideas), whereas academic tests tend to focus more on whether students can meet predetermined task expectations (e.g., accurately solving a problem in mathematics). Figure 2 provides a visual representation of where creativity and academic achievement tests tend to place their emphasis.

As depicted in Figure 2, these areas of emphasis map onto the conceptual definition of creativity ($C = O \times TC$), with creativity tests tending to focus on the originality (O) aspect of creativity and measures of academic achievement tending to focus on meeting predetermined task constraints (TC). The area of empirical overlap between these measures is therefore restricted to the narrow intersection between O and TC. We therefore might expect that the empirical relationship between creativity and academic achievement is constrained by the types of measures used to assess these

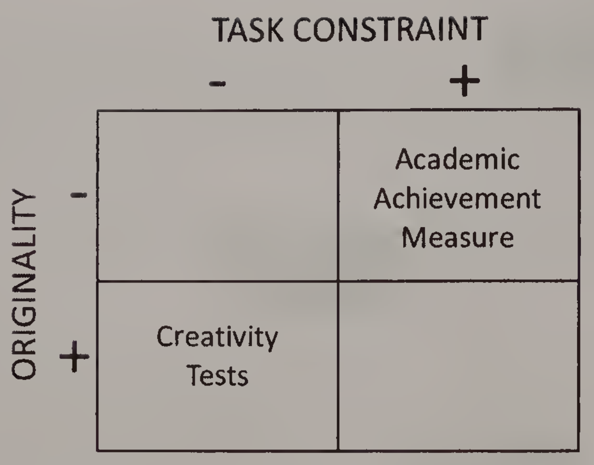


Figure 2. A visual representation of where creativity and academic achievement tests place emphasis.

constructs. It is unclear, however, how various types of measurement used in previous studies affect the average relationship between creativity and academic achievement. We therefore endeavor to shed light on this moderating factor.

Education Stage

Education stage is another potentially moderating factor in the relationship between creativity and academic achievement. Some of the earliest creativity research in classrooms was conducted by Torrance (1968), who documented what he called a fourth-grade slump (i.e., declines in creativity in the transition from third to fourth grade). Since that time, researchers have demonstrated variability in creativity scores across stages of education. The relationship between the imagination of children starting school and their achievements, for instance, hardly exists at all, $r = .02$ (Karwowski & Dziedziewicz, 2012). However, as early as the fifth grade, that relationship has been found to be more substantial, $r = .23$ (Jankowska, Gajda, & Karwowski, 2015; Karwowski, 2015). In yet other studies, the relationship between creativity and achievement in elementary school students has been found to range between $r = .08$ (Gajda, 2008) and $r = .39$ (Awamleh, Al Farah, & El-Zraigat, 2012). Variations have also been found in middle grades, $r = .18$ (Rindermann & Neubauer, 2004) and high school, ranging from $r = .12$ (Kim & Michael, 1995) to $r = .21$ (Karwowski, 2005).

Although education stage seems to moderate the relationship between creativity and achievement, there is no clear pattern or direction that can be expected from previous findings. As such, the present study aims to provide a more stable estimate of the influence of grade level on the relationship between creativity and academic achievement.

Decade

As with education stage, there is prior empirical work suggesting that decade may influence the relationship between creativity and academic achievement. Kim (2011) conducted one of the largest cross-sectional studies ($N = 279{,}599$) that examined the pattern of creativity scores over six time periods (from 1966–2008). Kim summarized her findings by stating that creativity scores, measured by the TTCT, are "declining overtime among Americans of all ages, especially kindergarten through third grade, the decline is steady and persistent, from 1990 to present, and ranges across various components tested by the TTCT" (p. 293). When taking the full range of decades into account, the patterns demonstrate more variability (including periods of gain, stagnation, and decline). Moreover, the changes from one sampled time period to the next are often (but not always) statistically significant, and the magnitude of the effect varies from small to large (depending on the particular component of the TTCT examined and the time period tested).

Consequently, we expect that time period likely will have some influence on the relationship between creativity and academic achievement, but it is difficult to predict the direction or magnitude of that difference. Our analysis will, however, allow us to examine whether studies conducted across different time periods moderate the relationship between creativity and academic achievement.

Culture

Finally, we expect culture to play a moderating role in the relationship between creativity and academic achievement. The direction and magnitude of that difference, however, are once again difficult to predict. Researchers have noted that conceptualizations of creativity can and do differ across cultures (Kaufman & Sternberg, 2006; Rudowicz, 2003). Given that there is so much variation within cultures (Freund & Holling, 2008; Gralewski & Karwowski, 2012), it is difficult to untangle the within variation from the between variation in previous work. As such, the present study aims to help clarify whether and to what extent culture moderates the relationship between creativity and academic achievement.

Method

Search Strategies

We followed a three-step procedure to select the studies included in our meta-analysis. The first step was a review of articles and research papers in English. We searched EBSCO, PsycExtra, Academic Search Complete, PsycInfo, PsycArticles, and ERIC databases and used the resources of JSTOR, Science Direct, SAGE Journals, Taylor & Francis, and ProQuest. In the next step, we analyzed book publications using three electronic libraries: Wiley Online Library and Questia, as well as Google Books.

We used the following search parameters to collect articles (keywords, abstracts, titles, and full text): *academic achievement** or *school grades** or *school achievement** or *scholastic achievement** or *grade point average* and *creative ability** or *creativity** or *divergent thinking**. Finally, in the third step of our search procedure, we explored whether any additional studies could be found by conducting a review of Polish-language periodicals devoted to psychology and education. We chose Polish-language periodicals because the first two authors had access to this literature and are fluent in the language.

Inclusion and Exclusion Criteria

Our search yielded a total of 148 studies. We then applied the following selection criteria to those studies. First, we only consid-

ered studies that presented a quantitative measure of the strength of the relationship between creativity and academic achievement, even if the relationship between creativity and academic achievement was not the primary goal of the study. A total of 18 studies[2] did not meet this first selection criterion and were eliminated from the analysis.

Next, we only included studies if they used more objective measures of creativity (e.g., TTCT) or self-report scales that demonstrated adequate reliability, such as measures of creative personality (e.g., Naderi et al., 2009) or creative self-confidence beliefs (e.g., Skager et al., 1967). This resulted in the elimination of four studies.[3] With respect to academic achievement, we included studies that used GPA (e.g., Chamorro-Premuzic, 2006), external examinations (e.g., Tan et al., 2014), and achievement tests created by researchers for the purpose of their study (e.g., Dobrołowicz, 2002; Sethi, 2012). This resulted in the elimination of one study that used students' self-assessments of academic achievement (Kaltsounis, 1974).

We also excluded two studies that used data presented in other publications, one study that used data previously published by a different author, and two studies that used multilevel models. The two studies that used multilevel models were excluded because they provided unstandardized regression coefficients that were inflated by the control of nesting students into classes and schools. Although β values are sometimes translated into *r* values (Peterson & Brown, 2005), there is no widely accepted or robust procedure for translating coefficients from multilevel models into standardized effect size for use in meta-analysis.

A total of 120 of the original 148 studies met our selection criteria and were included in the analysis. Taken together, the included studies had 782 effects with over 50,000 participants (N = 52,578). Participants had a mean age of 13.8 years (SD = 2.43) and attended elementary, middle, and high schools as well as colleges or universities. The studies were conducted between 1962 and 2015, in various countries (including the United States, European countries, Asia, and Africa). Table 1 provides a detailed overview of the studies included in the meta-analysis.

Coding Procedures

The first two authors independently coded each article for relevant information, including sample size, sample selection, effect size, and information necessary for the moderator analyses (i.e., measures of creativity and academic achievement, participants' age and stage of education, date and location of publication). Next, we reviewed the coded data and articles, as well as discussed and resolved any discrepancies to help eliminate errors in coding.

Moderators

For each study included in our analysis, we coded for the key moderators of interest. With respect to type of measurement, we coded the type of creativity measure used in the study (i.e., creative ability test or self-report questionnaire). We distinguished between different types of creativity tests, including tests based on Guilford's theory, the TCT-DP by Urban and Jellen, the TTCT, and other instruments (e.g., Remote Associates Test; Mednick, 1963). We also coded the different dimensions of creative ability measured by tests used in the studies (i.e., overall indices of creative ability, fluency, flexibility, originality of thinking, and elaboration). With respect to academic achievement, we coded for how achievement was measured (i.e., GPA or achievement test) and type of achievement measured (i.e., humanities, science, overall performance, sports).

Finally, we coded (a) education stage (i.e., elementary, middle school, high school, college/university), (b) study year (i.e., the year the study was conducted), and location (i.e., the country or continent where the study was conducted). We also included two dichotomously coded control variables that might influence the relationship between creativity and academic achievement. Those control variables included (a) goal of the study (i.e., primary purpose was examining the relationship between creativity and academic achievement vs. another goal) and (b) publication status (i.e., published or unpublished study).

Statistical Methods

When possible, we computed effect size using the values of correlation coefficients (r) and sample size (N). In a few studies, however, we converted the effect value provided (e.g., β, F, or χ^2) to the value of the r correlation coefficient. To analyze main effects, we used multilevel meta-analysis (Cheung, 2014, 2015; Konstantopoulos, 2011; Lebuda, Zabelina, & Karwowski, 2015), because individual correlations were clustered within studies. We carried out a three-level meta-analysis. Level 1 related to the participants in individual studies, Level 2 to interdependent effects within independent studies, and Level 3 to the studies themselves.

Three-level meta-analysis made it possible to obtain robust estimates of effect size, specifically unbiased estimates of standard errors, Level 2 (within-study) variance, and Level 3 (between-study) variance. Three-level meta-analysis was required because averaging the effects of individual studies would have significantly weakened the power of the entire analysis (we had 782 effects, but these were drawn from 120 studies) and would not have allowed us to estimate the influence of various moderators (as these were attributed to specific effects rather than studies).

---

[2] Those 18 studies focused on analyzing the theory of positive disintegration (Gallagher, 1985); the effectiveness of training influences (Blumen-Pardo, 2002; Cheung, Roskams, & Fisher, 2006; Malekian & Fathi, 2012; Yorke-Viney, 2007); the analysis of success in teaching (Hodder, 1972); the analysis of the relationship of sibling structure with creativity, intelligence, and academic achievement (Cicirelli, 1967); investigating the predictors of entrepreneurship (Farzaneh et al., 2010); seeking various predictors of academic achievement (Childs, 1978; Muhich, 1972; Owen, Feldhusen, & Thurston, 1970; Richards & Casey, 1975; Yamamoto, 1964); analyzing the relationship between parenting style, perfectionism, and creativity in talented individuals with high academic achievement (Miller, Lambert, & Speirs Neumeister, 2012); the teacher's perception of creativity, intelligence, and academic achievement (Mayfield, 1979); the measurement of creativity, intelligence, and academic achievement (Eisenman, Platt, & Darbes, 1968); analyzing the reliability and validity of ideational originality (Runco & Albert, 1985); and creativity in exact sciences (Son, 2009).

[3] Those four studies included a single self-report question from a questionnaire as a measure of creativity (Unal & Demir, 2009), judges' rating of participants with low-reliability products (Hasirci & Demirkan, 2007; Priest, 2006), and a questionnaire completed by the teacher concerning students' creativity level (Baltzer, 1988).

Table 1
The Subjects of the Studies Included in the Meta-Analysis

No.	Study	Aim of the study	Type of publication	Creativity measure	Achievement measure	*N*	*r*	No. of effects	School	Country	Year
1	Cline, Richards, & Abe, 1962	The reliability of creativity test batteries	Journal article	Guilford Tests of Divergent Thinking	GPA	161	.32	12	high school	United States	1962
2	Mednick, 1963	The creativity of psychology students	Journal article	Creativity Rating Scale Remote Associations Test	GPA	21	−.03	2	college/university	United States	1963
3	Cline, Richards, & Needham, 1963	Creativity and achievement tests in school education	Journal article	Guilford Tests of Divergent Thinking	GPA	119	.30	14	high school	United States	1963
4	DeBoer, 1964	The relationship of creativity with intelligence and academic achievement	Unpublished dissertation	Getzels, Jackson Creative Thinking Tasks	The Iowa Test of Basic Skills	277	.51	5	elementary, middle and high school	United States	1964
5	Karnes, Zehrbach, Studley, & Wright, 1965	The intellectual functioning of children with high potential from culturally threatened communities	Unpublished study	TTCT	SAT	203	.32	4	elementary	United States	1963
6	Ohnmacht, 1966	Academic achievement, anxiety, and creative thinking	Journal article	Getzels and Jackson Creative Thinking Tasks: 2. Uses for Things	SAT	204	.56	18	elementary	United States	1966
7	Yamamoto & Chimbidis, 1966	Academic achievement, intelligence, and creative thinking	Journal article	Minnesota Tests of Creative Thinking	SAT	790	.22	9	elementary	United States	1966
8	Bentley, 1966	Creativity and academic achievement	Journal article	Test of Imagination The-Ask-and-Guess Test	GPA	75	.34	1	college/university	United States	1966
9	Klein, Skager, & Erlebacher, 1966	Measures of artistic creativity and flexibility of thinking	Journal article	Hidden Figures Thurstone Complex Space Pitcher	GPA	92	.02	3	college/university	United States	1966
10	Dowd, 1966	The relationship between creative thinking and academic success	Unpublished study	Guilford Tests of Divergent Thinking	Scholastic Aptitude Test	417	.16	18	college/university	United States	1966
11	Yamamoto, 1967	Creativity and unpredictability as related to academic achievement	Journal article	Minnesota Tests of Creative Thinking	Iowa Tests of Educational Development	159	.04	18	high school	United States	1967
12	Skager, Klein, & Schultz, 1967	Predicting academic and artistic achievement	Journal article	Self-Ratings of Creativity Symbol Production Task Controlled Associations Task Alternate Uses Task	GPA	357	.27	7	college/university	United States	1967
13	Graves, Ingersoll, & Evans, 1967	The creative medicine student: a descriptive study	Journal article	Opinion, Attitude and Interest Survey	Medical College Admissions Test	150	.03	5	college/university	United States	1967
14	Mednick & Andrews, 1967	Creative thinking and the level of intelligence	Journal article	Remote Associations Test	Scholastic Aptitude Test	1,211	.32	2	college/university	United States	1967

Table 1 (*continued*)

No.	Study	Aim of the study	Type of publication	Creativity measure	Achievement measure	N	r	No. of effects	School	Country	Year
15	Hetrick, Lilly, & Merrifield, 1968	Figural creativity, intelligence, and children's personality	Journal article	Gulamerian Flexibility and Originality tests	Iowa Achievement Test GPA	196	.06	4	elementary	United States	1968
16	Wade, 1968	Differences between intelligence and creativity: suppositions concerning the role of the environment	Journal article	Guilford Tests of Divergent Thinking	GPA	105	.35	3	high school	United States	1968
17	Tibbetts, 1968	Socioeconomic status, race, gender, IQ, age, and GPA among adolescents	Unpublished dissertation	Getzels and Jackson Creativity Test	GPA	258	.42	2	high school	United States	1968
18	Bowers, 1969	Interactions between creativity and intelligence	Journal article	TTCT, verbal & figural	GPA Iowa Tests of Educational Development	278	.26	40	middle-school	United States	1969
19	Anderson, White, & Stevens, 1969	Creativity, intelligence, students' achievement, and behavior	Journal article	The Utility Test The Apparatus Test The Plot Titles Test	GPA Language Test CAT	197	−.03	2	high school	United States	1969
20	Kanderian, 1969	The relationship between academic achievement, creativity, and intelligence	Unpublished dissertation	Divergent Thinking Test	GPA	304	.05	3	elementary	Iraq	1969
21	Gluskinos, 1971	Types of students' creativity and their relationship with academic grades	Journal article	AC Test for Creative Ability The Purdue Creativity Test Guilford's Divergent production	GPA	314	.02	11	college/university	United States	1971
22	Feldhusen, Treffinger, Van Mondfrans, & Ferris, 1971	The relationship between school grades and the level of creative thinking	Journal article	TTCT, verbal & figural	GPA	356	.11	112	elementary and high school	United States	1971
23	Wagner, 1972	Time-bound and non-time-bound measures of creativity and their correlates	Unpublished dissertation	Wallach and Wing Test	GPA	131	.13	16	high school	United States	1972
24	Stallings & Gillmore, 1972	The relationship between nonverbal creativity and grades	Unpublished study	TTCT, figural	GPA ACT	292	.06	8	college/university	United States	1972
25	Kaltsounis & Stephens, 1973	Arithmetical achievement and creativity	Journal article	TCT, verbal & figural	Seeing Through Arithmetic Tests	196	.22	5	elementary	United States	1973
26	Porter, 1974	Race, socialization, and mobility in education and early professional achievement	Journal article	Getzels and Jackson Unusual Uses Creativity Test TTCT	GPA	15,326	.22	2	high school	United States	1974
27	Joesting, 1975	Creativity and school grades, school activities, and students' attendance	Journal article	Torrance's Creative Motivation Inventory Khatena-Torrance What Kind of Person Are You?	Achievement Test GPA	133	.16	10	college/university	United States	1975

(*table continues*)

Table 1 (*continued*)

No.	Study	Aim of the study	Type of publication	Creativity measure	Achievement measure	N	r	No. of effects	School	Country	Year
28	Marjoribanks, 1976	Regression analyses: academic achievement, intelligence, and creativity	Journal article	Adaptation of Torrance and Guilford Tests of Divergent Thinking	GPA	450	.41	18	middle-school	United States	1976
29	Rodriguez, 1980	The influence of creativity training on academic achievement and creative thinking	Unpublished dissertation	Divergent Thinking–Structure of Intellect Learning Abilities Test (SOILAT)	Comprehensive Test of Basic Skills	188	−.05	4	elementary	United States	1980
30	Wódz, 1981	Students' creative talents and their academic achievement	Journal article	Wallach Kogan Creativity Test	GPA	60	.24	1	elementary	Poland	1981
31	Stiles, 1982	Correlates of creativity among students	Unpublished dissertation	TTCT, verbal & figural	GPA	68	.14	8	college/university	United States	1982
32	Jackson, 1982	The relationship of cognitive abilities with intelligence and academic achievement	Unpublished dissertation	TTCT	Comprehensive Test of Basic Skills Prescriptive Reading Test Prescriptive Mathematics Test	76	.30	4	elementary	United States	1982
33	Westcott, 1983	Signature humor as an alternative in the identification of creative and talented students	Unpublished dissertation	TTCT, figural	The Iowa Tests of Basic Skills GPA	62	.43	6	high school	United States	1983
34	Toll, 1985	The relationship between Guilford's tests and reading comprehension	Unpublished dissertation	New Uses Test Seeing Different Meanings Test Judging Object Adaptations Test	Gates-Mac Ginitie Reading Test scores	105	.24	3	elementary	United States	1985
35	Janes, 1988	The relationship between creativity test scores and receiving awards and school grades	Unpublished dissertation	TTCT, figural	GPA	209	.09	4	high school	United States	1988
36	Orieux, 1989	Correlates of creative abilities and achievement in high schools	Unpublished dissertation	The Creative Activities Checklist Guilford Tests of Divergent Thinking	GPA	157	.40	4	high school	United States	1989
37	McCabe, 1991	The influence of creativity and unpredictability on academic achievement	Journal article	TTCT, verbal & figural	GPA	210	.20	21	middle-school	United States	1991
38	Guastello, Bzdawka, Guastello, & Rieke, 1992	Cognitive abilities and creative behavior	Journal article	CAB 5 The Remote Consequences Test	GPA	144	.18	5	college/university	United States	1992
39	Popov, 1992	Creativity and reading comprehension	Journal article	Guilford Tests of Divergent Thinking	Reading Comprehension Test	63	.38	10	college/university	Russia	1992
40	Kim & Michael, 1995	The relationship of creativity with academic achievement and the preferred style of thinking	Journal article	TTCT, verbal & figural	GPA	193	.12	5	high school	Korea	1995

Table 1 (*continued*)

No.	Study	Aim of the study	Type of publication	Creativity measure	Achievement measure	N	r	No. of effects	School	Country	Year
41	Auzmendi, Villa, & Abedi, 1996	The validity and reliability of an instrument for measuring creativity	Journal article	Abedi-Schumacher Creativity Test	GPA	2,264	.14	24	no data	Spain	1996
42	Johns & Morse, 1997	Divergent thinking	Journal article	The Alternative Uses Test (Based on Guilford and Wallach, Kogan tests)	GPA	79	.28	3	college/university	United States	1997
43	Simpson, 1998	The relationship of academic achievement with intelligence, creativity, motivation, and gender identification	Unpublished dissertation	TTCT, figural	The Texas Learning Index The Texas Assessment of Academic Skills	71	.14	2	elementary	United States	1998
44	Niaz, Núñez, & Pineda, 2000	Students' academic achievement as an effect of mental abilities, cognitive style, fixation vs. mobility dimensions, and creativity	Journal article	TTCT, figural	GPA	141	.16	32	high school	Venezuela	2000
45	Diakidoy & Constantinou, 2001	Creativity in physics: the fluency of answers and the specificity of tasks	Journal article	Three Open Tasks	GPA	54	.11	4	college/university	Greece	2000
46	Gollmar, 2000	Attention-deficit hyperactivity disorder, creativity, and cognitive styles: interactions and effect on school success	Unpublished dissertation	TTCT	GPA Iowa Test of Basic Skills	179	.08	20	elementary	United States	2000
47	Clapham, 2001	The effect of information manipulation and exposure on divergent thinking	Journal article	TTCT	GPA	148	.14	1	college/university	United States	2001
48	Dobrołowicz, 2002	Students' creativity and their academic achievement	Book	Modified Guilford Tests of Divergent Thinking	School Achievement Tests	445	.28	9	elementary	Poland	2002
49	Grzelak, 2003	Creative and imitative attitudes and academic achievement in third-grade high school students	Unpublished thesis	Creative Behavior Questionnaire	GPA	200	.06	1	high school	Poland	2003
50	Chamorro-Premuzic & Furnham, 2003	Personality and the prediction of academic achievement	Journal article	The Barron-Welsh Art Scale	GPA	70	.07	2	college/university	Great Britain	2003
51	Cheung, Rudowicz, Yue, & Kwan, 2003	The influence of the year of study on university students' creativity	Journal article	Alternate Uses Test Guilford	GPA	859	.09	6	college/university	China	2003

(*table continues*)

Table 1 (*continued*)

No.	Study	Aim of the study	Type of publication	Creativity measure	Achievement measure	N	r	No. of effects	School	Country	Year
52	Zhang, 2004	School performance prediction styles	Journal article	Sternberg Triarchic Abilities Test	GPA	250	.28	16	middle-school	Hong Kong	2004
53	Yeh, 2004	Academic achievement, creativity, and mind-mapping ability	Journal article	The Alternative Uses Test by WU	GPA	69	.66	3	middle-school	Taiwan	2004
54	Rindermann & Neubauer, 2004	Processing speed, intelligence, creativity, and school performance	Journal article	Verbaler Kreativity-Test Verwendungs Test	GPA	271	.18	8	middle and high school	Germany	2004
55	Clapham, 2004	The reliability of the TTCT test battery and self-report	Journal article	How Creative Are You? How Do You Think?	SAT ACT	145	.18	8	college/university	United States	2004
56	Karwowski, 2005	The school performance and creativity of students perceived by teachers as different in terms of abilities	Journal article	Urban, Jellen TCT-DP	GPA	194	.20	1	high school	Poland	2005
57	Uszyńska-Jarmoc, 2005	Types of thinking and the academic achievement of 7-year-olds	Journal article	Urban, Jellen TCT-DP	GPA	167	.10	2	elementary	Poland	2005
58	Al-Dhobaiban, 2005	The relationship between self-regulation and creativity	Unpublished dissertation	Test Your Creativity Level Scale Khatena Torrance Creativity Perception Inventory	GPA	219	.15	2	college/university	United States	2005
59	Stankiewicz, 2006	The talented student in the teacher's eyes	Unpublished thesis	Urban, Jellen TCT-DP	GPA	79	.19	1	high school	Poland	2006
60	Chamorro-Premuzic, 2006	Creativity and conscientiousness: which of the variables is a better predictor of a student's achievement?	Journal article	Christensen Alternate Uses Test	GPA	307	.19	3	college/university	Great Britain	2006
61	Sternberg, 2006	Enhancing the SAT through assessments of analytical, practical, and creative skills	Journal article	Sternberg Triarchic Abilities Test	GPA	777	.40	2	college/university	United States	2006
62	Arseniuk, 2007	Creativity differences and disability	Unpublished thesis	Urban, Jellen TCT-DP	GPA	130	.24	1	high school	Poland	2007
63	Freund, Holling, & Preckel, 2007	The relationship between cognitive abilities and academic achievement	Journal article	Berlin Structure of Intelligence Test for Youth: Assessment of Talent and Giftedness	GPA	1,135	.38	3	middle-school	Germany	2007
64	Wang, 2007	Teaching, learning, and creativity in Taiwan and the United States	Unpublished dissertation	Abbreviated Torrance Test for Adults	Entrance Examinations CBEST	216	.27	3	college/university	United States	2007

Table 1 (*continued*)

No.	Study	Aim of the study	Type of publication	Creativity measure	Achievement measure	*N*	*r*	No. of effects	School	Country	Year
65	Wasil, 2008	The student's creativity as related to his or her grades and position in the teacher's eyes	Unpublished thesis	Urban, Jellen TCT-DP	GPA	30	.20	2	high school	Poland	2008
66	Dąbrowska, 2008	Creative abilities, school performance, and motivation type as related to interpersonal attractiveness	Unpublished thesis	Urban, Jellen TCT-DP	GPA	75	.10	1	middle-school	Poland	2008
67	Gajda, 2008	The accuracy of teachers' nominations and the actual level of students' creative abilities	Unpublished thesis	Urban, Jellen TCT-DP	GPA	90	.08	1	elementary	Poland	2008
68	Hirsh & Peterson, 2008	The prediction of creativity using an instrument connected with personality traits	Journal article	Creative Achievement Questionnaire	GPA	98	.01	1	college/university	Canada	2008
69	Palaniappan & Persekutuan, 2008	The influence of intelligence on the relationship between creativity and academic achievement	Journal article	TTCT, figural	GPA	72	.17	8	high school	Malaysia	2008
70	Sternberg, 2007	Improving school-related skills by diversifying the goals to be achieved	Journal article	Sternberg Triarchic Abilities Test	GPA Scholastic Assessment Test (Writing and Math) SAT	793	.50	4	college/university	United States	2008
71	Silvia, 2008	Creativity and intelligence: a latent variable analysis by Wallach and Kogan	Journal article	Guilford Tests of Divergent Thinking	School and College Ability Test Sequential Tests of Educational Progress	151	.25	1	elementary	United States	2008
72	Zielińska, 2009	The significant others of young people with different levels of creativity	Unpublished thesis	Urban, Jellen TCT-DP	GPA	115	.07	1	high school	Poland	2009
73	Tkaczyk, 2009	Conflict resolution styles and selected aspects of middle school students' creativity	Unpublished thesis	Urban, Jellen TCT-DP	GPA	209	.18	1	middle-school	Poland	2009
74	Karwowski, Lebuda, & Wiśniewska, 2008–2009	Creative abilities and styles as predictors of school success	Journal article	Urban, Jellen TCT-DP	GPA	1,316	.11	2	middle and high school	Poland	2009
75	Naderi, Abdullah, Aizan, Sharir, & Kumar, 2009	Creativity, age, and gender as predictors of students' academic achievement	Journal article	Khatena-Torrance Creative Perception Inventory	GPA	153	.16	1	college/university	Iran	2009

(*table continues*)

Table 1 (*continued*)

No.	Study	Aim of the study	Type of publication	Creativity measure	Achievement measure	N	r	No. of effects	School	Country	Year
76	Karwowski, Ciak, & Grubek, 2009	The accuracy of teachers' nominations in recognizing students' creativity	Book chapter	Urban, Jellen TCT-DP Creative Behavior Questionnaire	GPA	94	.12	3	high school	Poland	2009
77	Mann, 2009	The identification of creative potential in middle school students	Journal article	Creative Abilities in Mathematics Test	Connecticut Mastery Tests	89	.48	1	middle-school	United States	2009
78	Bierly, Kolodinsky, & Charette, 2009	The relationship between creativity and ethical ideologies	Journal article	Creativity Scale based on Kirton and Robinson theories	GPA	899	.04	1	college/university	United States	2009
79	Olatoye, Akintunde, & Yakasai, 2010	Emotional intelligence, creativity, and students' academic achievement	Journal article	Nicola Holt Creativity Test: The Creative Cognition Inventory	GPA	235	.00	1	college/university	Nigeria	2010
80	Day, Hanson, Maltby, Proctor, & Wood, 2010	Hope as a predictor of academic personality and academic achievement	Journal article	Gilford Tests of Divergent Thinking	GPA	129	.31	1	college/university	Great Britain	2010
81	Vijetha & Jangaiah, 2010	Intelligence, creative thinking, and academic achievement in children with hearing disorders	Journal article	Baqer Mehdi Nonverbal Test of Creativity	GPA	50	−.07	4	elementary	India	2010
82	Kousoulas, 2010	The interrelations between creative behavior, divergent thinking, and knowledge during students' creative expression at the time of learning	Journal article	Two Open Tasks	Achievement Test GPA	115	.26	12	elementary	Greece	2010
83	Sak & Maker, 2006	The developmental differentiation of mathematical creative thinking in children depending on age, education, and the level of knowledge	Journal article	DISCOVER (based on Guilford Divergent Thinking Tests)	Mathematical Knowledge DISCOVER	297	.38	2	elementary	United States	2010
84	Dhatrak & Wanjari, 2011	The relations between school attitude, creativity, and academic achievement	Journal article	Verbal Test of Creative Thinking Baqer Mehdi	GPA	500	.05	1	high school	India	2011
85	Pishghadam, Khodadady, & Zabihi, 2011	Creativity and learning foreign languages	Journal article	Arjomand Creativity Questionnaire	GPA	272	.36	1	college/university	Iran	2011
86	Dollinger, 2011	Recruitment tests and creativity	Journal article	Hocevar Creative Behavior Inventory	ACT	433	.17	4	college/university	United States	2011

Table 1 (*continued*)

No.	Study	Aim of the study	Type of publication	Creativity measure	Achievement measure	N	r	No. of effects	School	Country	Year
87	Dziedziewicz & Gajda, 2011	Creativity and academic achievement in middle school	Unpublished study	Urban, Jellen TCT-DP	GPA 3-grade and 6-grade Achievement Tests	362	.29	2	middle-school	Poland	2011
88	Petrulyte, 2011	Creativity and academic achievement in students of arts	Journal article	TTCT Asmenybes Kurybiskumo Klausimynas	GPA	140	.24	1	high school	Lithuania	2011
89	Vock, Preckel, & Holling, 2011	Intellectual abilities and academic achievement: analysis of mediators	Journal article	Berlin Structure of Intelligence Test for Youth: Assessment of Talent and Giftedness	GPA	1,135	.38	3	high school	Germany	2011
90	Ofili, 2011	Do multicultural experiences and biculturalism enhance students' creativity?	Unpublished thesis	Creative Achievement Questionnaire	GPA	122	.16	1	college/university	United States	2011
91	Hansenne & Legrand, 2012	Creativity, emotional intelligence, and academic achievement in students	Journal article	TTCT	GPA	73	.20	14	elementary	Belgium	2012
92	Piórkowska, 2012	The dynamics of strategic behaviors in learning and their associations with creative abilities	Unpublished thesis	Urban, Jellen TCT-DP	GPA	107	.17	1	high school	Poland	2012
93	Scthi, 2012	Investigating performance in math in relation to the creativity of high school students	Journal article	Verbal Test of Creative Thinking Baqer Mehdi	Mathematics Achievement Test for 9th Class	700	.26	3	high school	India	2012
94	Anwar, Aness, Khizar, Naseer, & Mubammad, 2012	The relationship between creative thinking and students' academic achievement	Journal article	TTCT	GPA	256	.46	4	high school	Pakistan	2012
95	Awamleh, Al Farah, & El-Zraigat, 2012	Creativity level measured using the TTCT	Journal article	TTCT	GPA	63	.39	4	elementary	Jordan	2012
96	Putwain, Kearsley, & Symes, 2012	Creative self-awareness as related to literary achievement and motivation	Journal article	Abedi Test of Creativity	Achievement Test	122	.23	4	middle-school	Great Britain	2012
97	Karwowski & Dziedziewicz, 2012	Correlations between creative imagination and skills at the beginning of school	Book	Creative Imagery Abilities Test	Test of Skills at School Beginning	1,096	.02	12	nursery and elementary	Poland	2012
98	Walia, 2012	The relationship between achievement and mathematical creativity	Journal article	Balka Creative Ability in Mathematics Test	Sessional Assessment in Mathematics Test	180	.66	3	middle-school	India	2012

(*table continues*)

Table 1 (*continued*)

No.	Study	Aim of the study	Type of publication	Creativity measure	Achievement measure	N	r	No. of effects	School	Country	Year
99	Tatlah, Aslam, Ali, & Iqbal, 2012	The role of intelligence and creativity in academic achievement	Journal article	Nicola Holt Creativity Test: The Creative Cognition Inventory	GPA	235	.03	1	college/university	Pakistan	2012
100	Dai et al., 2012	The influence of social environment and education conditions on creativity in adolescence	Journal article	Four-item creativity test	GPA	229	.35	2	middle-school	United States	2012
101	Ibrahim, 2012	The relationship between creativity, engineering, knowledge, and design	Unpublished dissertation	TTCT	GPA	85	.03	7	college/university	United States	2012
102	Gajda, 2013	Creativity and academic achievement, elementary schools	Unpublished dissertation	Urban, Jellen TCT-DP	GPA	384	.16	1	elementary	Poland	2013
103	Karwowski, 2013a	Creativity and academic achievement	Unpublished study	Urban, Jellen TCT-DP	GPA	392	.14	1	no data	Poland	2013
104	Karwowski, 2013b	Creativity and academic achievement	Unpublished study	Urban, Jellen TCT-DP	GPA	113	.03	1	no data	Poland	2013
105	Dziedziewicz, 2013	Creativity and learning styles	Unpublished study	Creative Imagery Abilities Test	GPA	238	.14	1	college/university	Poland	2013
106	Mitrofanow, 2013	Creative imagination and academic achievement	Unpublished thesis	Creative Imagination Test by Kujawski	GPA	140	.12	3	college/university	Poland	2013
107	Ritchie, Luciano, Hansell, Wright, & Bates, 2013	The relationship between reading-related skills and creativity	Journal article	Queensland Core Skills Test: "create and present" scores	Reading and Spelling Test Core	855	.32	1	no data	Australia	2013
108	Gralewski & Karwowski, 2013	Students' gender and the accuracy of teachers' creativity nominations	Journal article	Creative Activity in Science Creative Activity in Arts Urban Jellen TCT-DP Creative Behavior Questionnaire	GPA	589	.05	5	high school	Poland	2013
109	Lovelace & Hunter, 2013	The influence of charismatic ideological and pragmatic leaders on the creative process and the products of creativity	Journal article	Divergent Thinking Test of Fluency (Baer, 1993)	SAT	336	.03	4	college/university	United States	2013
110	Matthew & Stemler, 2013	The assessment of the flexibility of thinking based on a new word recognition test	Journal article	Abbreviated Torrance Test for Adults	SAT	299	.17	5	college/university	United States	2013
111	Tan, Mourgues, Bolden, & Grigorenko, 2014	Two methods of scoring creativity in the Aurora test	Journal article	Cartoon Numbers is a subtest of the Aurora Battery	Key Stage 2 Exam Tests MidYIS (Middle Years Information System)	205	.66	2	elementary	Great Britain	2013

Table 1 (*continued*)

No.	Study	Aim of the study	Type of publication	Creativity measure	Achievement measure	*N*	*r*	No. of effects	School	Country	Year
112	Zabelina, Condon, & Beeman, 2014	The relationship of psychopathology with creative achievement and creative thinking	Journal article	Abbreviated Torrance Test for Adults Creative Achievement Questionnaire	ACT SAT	100	.11	2	college/university	United States	2014
113	Chooi, Long, & Thompson, 2014	Sternberg Triarchic Abilities Test as an instrument for measuring intelligence	Journal article	Sternberg Triarchic Abilities Test	GPA	260	.37	3	college/university	United States	2014
114	Szumski & Karwowski, 2014	Investigating the influence of students with special educational needs on the functioning of their healthy peers	Unpublished study	Divergent Thinking Tests	Achievement Tests (Mathematics and Language)	1,611	.25	2	middle-school	Poland	2014
115	Karwowski & Szumski, 2014	Effects of intraclass comparisons	Unpublished study	Guilford Tests of Divergent Thinking	Achievement Tests (Mathematics and Language)	3,312	.24	4	middle-school	Poland	2014
116	Hunter & Cushenbery, 2015	Examining the role of disagreeableness in the sharing and utilization of original ideas	Journal article	Guilford Tests of Divergent Thinking	GPA Scholastic Assessment Test (Writing and Math) SAT	201	.06	6	college/university	United States	2015
117	Wallace & Russ, 2015	Pretended play, divergent thinking, and mathematical achievement	Journal article	Wallach, Kogan Test of Creative Thinking	AIMS Web Achievement Scores Test	24	.54	5	elementary	United States	2015
118	Pretz & Kaufman, in press	University admission criteria and applicants' creativity	Journal article	Creative Self-Efficacy Guilford Tests of Divergent Thinking	Scholastic Assessment Test (Writing and Math) SAT	363	.04	12	college/university	United States	2015
119	Jankowska, Gajda, & Karwowski, 2015	The determinants of education efficiency	Journal article	Creative Imagery Abilities Test	Mathematic Achievement Test Language Achievement Test	5,174	.21	4	elementary	Poland	2015
120	Roķe & Kālis, 2015	The relationship between academic achievement and creativity	Journal article	Urban, Jellen TCT-DP	GPA	180	.27	15	middle-school	Lithuania	2015

Note. GPA = grade point average; TTCT = Torrance Test of Creative Thinking; SAT = Stanford Achievement Tests; ACT = American College Testing; CAB 5 = Comprehensive Ability Battery; TCT-DP = Test of Creative Thinking–Drawing Production; CBEST = California Basic Educational Skills Test.

The multilevel meta-analysis was conducted using the metaSEM package (Cheung, 2014, 2015) in the R statistical environment (R Development Core Team, 2013). When analyzing the effect of publication bias, we also used the Comprehensive Meta-Analysis package (Biostat, 2008), the metafor package in R (Viechtbauer, 2010), and p-curve (Simonsohn, Nelson, & Simmons, 2014).

Results

We present the results of the meta-analysis in three steps. First, we present a general estimation of the effect size obtained in the multilevel model and in the random-effects model. Next, we analyze the potential influence of publication bias, which helps determine the robustness of the obtained effect size. Finally, in further multilevel models, we present the results of our moderator analyses.

Overall Effect

Table 2 presents the overall effect of the relationship between creativity and academic achievement. The obtained mean effect size was consistent with our expectations. More specifically, there was a positive and statistically significant, albeit modest, relationship: $r = .22$, 95% CI [.19, .24].[4] As expected, this effect was also heterogeneous, $Q(df = 781) = 9{,}481.65$, $p < .001$. Both within-study variance (between particular effects) and between-study variance were statistically significant, with most of the variance being between ($I^2 = .62$) rather than within studies ($I^2 = .30$).[5] Prior to examining the influence of moderators, however, we examined to what extent the obtained effect may be influenced by publication bias.

Publication Bias

We analyzed the robustness of the obtained effect size by examining whether it was influenced by publication bias. We used a four-step process that included both classic and more recent methods of analysis. First, we used a funnel plot (Duval & Tweedie, 2000) with several nonparametric techniques to estimate possible bias. We next used a p-curve analysis (Simonsohn et al., 2014) and then estimated the effect of using PET-PEESE[6] (Stanley & Doucouliagos, 2014). Finally, we compared effect sizes obtained in published versus unpublished studies.

An inspection of the funnel plot (see Figure 3) does not suggest asymmetry (i.e., correlations on one side of the funnel do not seem to be regularly suppressed by the effects on the other side). This pattern suggests a lack of publication bias (although such an interpretation is based more on a qualitative judgment, rather than strict statistical rules).

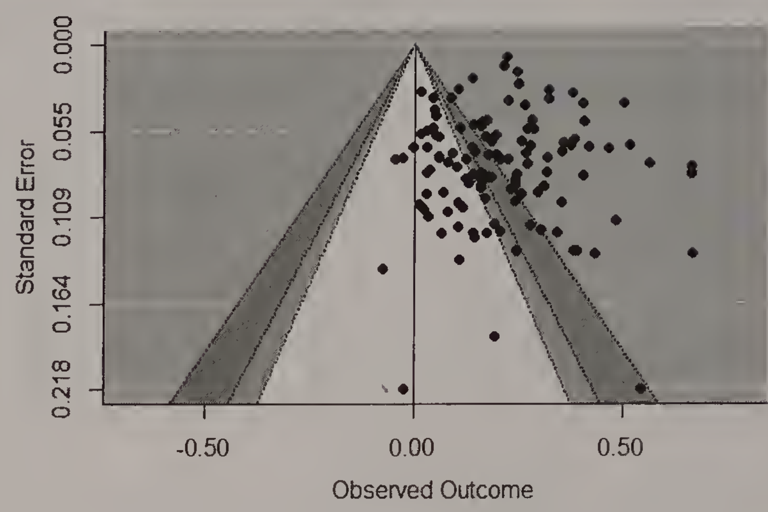


Figure 3. A funnel plot assessing the possible publication bias.

To assist with the interpretation of the funnel plot, researchers conducting meta-analyses often include statistical analysis. We used Egger's regression intercept test (Egger, Davey Smith, Schneider, & Minder, 1997). Based on the random effects model, assessing funnel plot asymmetry, and Begg and Mazumdar (1994) rank correlation test (nonsignificant $ps = .42$ and .30, respectively), we concluded there was no evidence of publication bias.

We next performed a p-curve analysis[7] (Simonsohn et al., 2014) to examine the credibility of the estimate using the online application available at http://www.p-curve.com/. The results of the p-curve analysis (see Figure 4) provided no evidence of a "file-

Table 2
Overall Effect Size Obtained Using Three-Level Meta-Analysis

Effects	Estimate	*SE*	95% CI *LL*	95% CI *UL*	*p*
Fixed effect					
Overall effect	.215	.015	.187	.244	<.001
Random effects					
Within-study variance	.010	.001	.008	.011	<.001
Between-study variance	.020	.003	.013	.026	<.001

Note. Number of studies = 120, number of effects = 782, total N = 52,578. CI = confidence interval; LL = lower limit; UL = upper limit.

[4] Robustness check performed using Comprehensive Meta-Analysis software (Biostat, 2008) on averaged effects for studies revealed the existence of an identical relationship. Due to high heterogeneity (Q = 892.61, $df = 119$, $p < .001$, $I^2 = 86.67\%$), we performed analyses using the random-effects model, in which we obtained a mean correlation of r = .22, 95% CI [.19, .24], and a high degree of heterogeneity, $\tau^2 = .015$, τ = .12.

[5] The relatively low within-study variance compared to between-study variance suggests that an equally good analytic choice could have been meta-analysis using the random-effects method on data aggregated to the level of individual studies. However, we chose multilevel analysis performed at the level of individual correlations (with correlation grouping in studies controlled for), because some of the possible moderators clearly had a within-study character (e.g., the operationalization of creative abilities as the fluency, flexibility, and originality of thinking).

[6] This method fits a meta-regression model predicting effect sizes in studies by their variances (the precision effect test, called PET) or their standard errors (the precision effect estimate with standard errors, called PEESE). If the intercept is statistically significant in the PET model, the PEESE model should be taken into account as the publication bias–free effect size.

[7] The p-curve analysis focuses only on statistically significant studies (i.e., all effects below significance level are excluded) and checks whether "just significant effects" (i.e., slightly lower than $p = .05$ or between p = .04 and $p = .05$) are not overrepresented in the analyzed studies. Such overrepresentation may be caused not only by publication bias but also by "cherry-picking," "p-hacking," or other questionable research practices (Simonsohn et al., 2014).

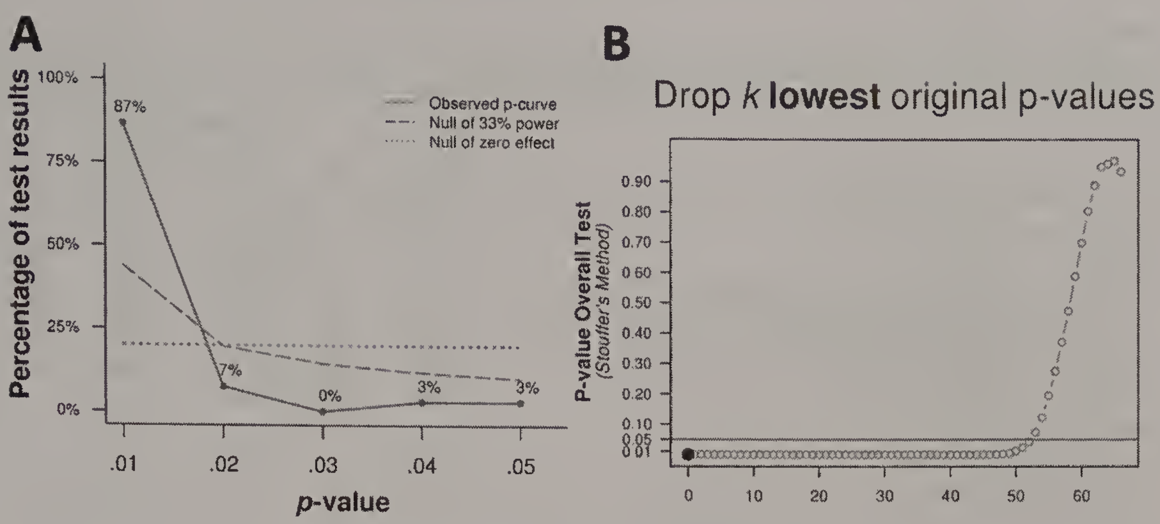


Figure 4. The p-curve analysis of publication bias. (A) The p-curve analysis and (B) its robustness. See the online article for the color version of this figure.

drawer effect" (i.e., most studies provided highly significant results) and there was also no overrepresentation of "just significant effects" (i.e., slightly lower than $p = .05$ or between $p = .04$ and $p = .05$; Figure 4A). Even more important, p-curve analysis demonstrated that the obtained effects were quite robust and insensitive to the exclusion of subsequent studies with the highest p values (Figure 4B).

Moreover, the continuous test for a right-skewed curve (i.e., examining whether studies contain evidential value) was statistically significant ($z = -30.78$, $p < .0001$), whereas testing for left-skewed studies (i.e., those that exhibit evidence of p-hacking) did not yield significant results ($p > .999$). Taken together, the results of our p-curve analysis provided further evidence that there was not an influence of publication bias.

The third step in the analysis of publication bias was the creation of a model based on the PET-PEESE method. Because the intercept obtained using PET was statistically significant ($r = .219$, $SE = .016$, $p < .001$), we adopted the intercept obtained based on PEESE as a measure of effect size not affected by publication bias, as recommended by Stanley and Doucouliagos (2014). The obtained effect was nearly the same as the results reported before (i.e., $r = .215$, 95% CI [.192, .239], $p < .001$), which also suggests no evidence of publication bias.

Finally, the results of comparing the effects obtained in published, $r = .23$, 95% CI [.20, .27], versus unpublished studies, $r = .19$, 95% CI [.15, .22], revealed a marginal difference in favor of published studies ($Q = 3.86$; $df = 1$, $p = .05$), but the similar size of the estimated effects and the overlapping confidence intervals make it legitimate to conclude that publication bias did not substantively influence our estimations.

Moderator Analysis

We analyzed the role of moderators in a sequence of multilevel regression models and used the measures of the baseline model's fit ($-2LL = -794.92$, $df = 3$) obtained in our analysis of the overall effect to compare models with moderators included as predictors. This approach allowed us to control for the mutual associations between predictors. At the end of this section, we also present estimations obtained using the analysis of variance (ANOVA) analog (Wilson, 2014), performed at the study level. Although not as statistically robust as multilevel models, the ANOVA analog analysis provides estimated effects for different groups of studies in a more convenient and easier to interpret fashion.

Types of measurement and study year. In the first step, we entered three moderators representing the measurement of creativity (0 = *self-report,* 1 = *test*), academic achievement (0 = *test,* 1 = *GPA*), and study year (grand centered). We also included two control variables in Step 1: research objective (0 = *other*, 1 = *creativity*—achievement) and publication status (0 = *unpublished*, 1 = *published*).

This model demonstrated better fit to the data ($-2LL = -813.61$, $df = 8$, $\Delta\text{-}2LL = 18.69$, $\Delta df = 5$, $p = .002$) compared to our baseline model. The results are presented in Table 3.

As displayed in Table 3, the predictors entered in the model explained 11% of between-study variance and 2.2% of within-study variance. The obtained effects were stronger when creativity was measured using tests compared to when it was measured using self-report scales, as well as stronger for academic achievement measured using standardized tests compared to using GPA. With respect to study year, there was no significant influence on the obtained effect size, suggesting that the correlations were stable across time (see Figure 5). Finally, the two control variables (i.e., research objective and publication status) were not significantly related to effect size.

In the second step, we removed nonsignificant predictors from the model (research objective, publication status, study year) and added variables specifying the location of study (with Europe as the reference value) and type of achievement measured (i.e., performance in the humanities, in sciences, and overall performance, with sport as the reference value). This model did not fit the data significantly better than the previous model ($-2LL = -819.37$, $df = 13$, $\Delta\text{-}2LL = 5.76$, $\Delta df = 5$, $p = .33$). Given that these additional moderators did not influence the obtained effects, our results indicate that the relationship between creativity and achievement was stable regardless

Table 3
Moderator Analysis: Types of Measurement and Study Year

Effects	Estimate	*SE*	95% CI *LL*	95% CI *UL*	*p*
Fixed effects					
Intercept	.119	.040	.041	.198	.003
Creativity measurement (0 = self-report, 1 = test)	.097	.028	.042	.153	.001
Academic achievement measurement (0 = test, 1 = GPA)	−.039	.018	−.074	−.004	.03
Study year (grand centered)	.0002	.001	−.001	.002	.76
Goal (0 = other, 1 = creativity × achievement)	−.003	.029	−.060	.054	.91
Published? (0 = no, 1 = yes)	.054	.030	−.005	.114	.07
Random effects					
Within-study variance	.009	.001	.008	.011	<.001
Between-study variance	.018	.003	.012	.023	<.001

Note. Number of studies = 120, number of effects = 782, total N = 52,578. CI = confidence interval; *LL* = lower limit; *UL* = upper limit; GPA = grade point average.

of location[8] where the study was conducted and regardless of domain of achievement examined. Moreover, given that these additional moderators were not significant, we do not provide detailed results of Step 2 of the analysis (but interested readers can find those results in the online supplemental material Table S1).

Education stage. The next step took into account the possibility of effects being influenced by the participants' education stage. We used a different model for examining this moderator because eight studies (and 154 correlations) used samples that combined participants from elementary and middle, elementary and high, or middle and high schools. Thus, we removed those eight studies from this step and conducted our analysis using a model that included a total of 628 effects from 112 studies (see Table 4).

The results of multilevel regression, using elementary school students as the reference category, indicated that the effect observed for middle school students was significantly higher than the effect for elementary students (B = 0.12, SE = 0.05; p = .015). The effect sizes obtained for high school and university/college students did not differ significantly from the effect obtained for elementary school students.

Aspects of creativity tests. Given that we found consistently stronger associations between creativity and academic achievement obtained in studies where creativity was measured using tests compared to self-report, we conducted a more focused analysis on studies that used creativity tests (i.e., 106 studies, 700 effects). The overall effect obtained only in those studies was r = .23, SE = .016, 95% CI [.20, .26], with a significant level of heterogeneity, $Q(df$ = 699) = 8,145.81, p < .001, situated mainly between studies, I^2 = .60, rather than within them, I^2 = .31, −2LL = −676.70, df = 3.

Therefore, in the next model, in addition to the method of measuring academic achievement, we included four more specific moderators in the group of creativity test predictors—namely, fluency, flexibility, originality of thinking, elaboration, and overall creative ability (e.g., the sum of TTCT or TCT-DP scores) and other measures (e.g., imagination as measured by Jankowska & Karwowski, 2015). We used a combination of overall indices of creative ability and other measures as the reference category for our analysis. This model did not fit the data better than the previously tested model (−2LL = −681.35, df = 8; Δ−2LL = 4.65, Δdf = 5; p = .46). Moreover, the various aspects of creative ability (fluency, flexibility, originality, elaboration) did not differ from the reference category in terms of the effect size generated (see the online supplemental material Table S3).

Next, we examined whether the verbal or figural characteristics of the creativity test resulted in different obtained effects. For this analysis, from the total pool of studies using creativity tests (106 studies, 700 effects), we excluded 16 studies whose authors did not provide separate results for verbal and figural tests (e.g., Anwar, Aness, Khizar, Naseer, & Muhammad, 2012; Porter, 1974; Zabelina, Condon, & Beeman, 2014). The observed effect was therefore estimated on a total of 90 studies. The results of this model are presented in Table 5.

As depicted in Table 5, the average effect size estimated on 617 correlations did not differ significantly from the overall effect previously reported: r = .228, SE = .017, 95% CI [.194, .262], p < .001; $Q(df$ = 616) = 7,520.86, p < .001; I^2 between studies = .595; I^2 within studies = .322 (−2LL = −577.52, df = 3). This model—which examined test type (0 = *figural*, 1 = *verbal*), in addition to the previously examined moderators and controls—fit the data better than the previously tested model (−2LL = −685.00, df = 15; Δ−2LL = 107.48, Δdf = 12; p < .001). Moreover, the results of this analysis indicate that verbal tests of creativity generated significantly higher effects than figural tests.

[8] An anonymous reviewer questioned our decision to include studies published in languages other than English (especially Polish but also Lithuanian). Specifically, the reviewer recommended that we exclude these studies as they may cause difficulty for those who want to replicate our study. Ultimately, we decided to keep these non-English studies in our analysis for three reasons. First, eliminating them would reduce the statistical power of our meta-analysis to 88 studies. Second, our additional analyses (see the online supplemental material Table S2) showed that although studies published in Polish and Lithuanian yielded significantly lower effect size, r = .14, 95% CI [.10, .18], than studies published in English, r = .24, 95% CI [.21, .27], this effect was caused by the fact that nonverbal tests were more often used in Poland and Lithuania, not by the country itself. When we controlled for the type of the test, the effect of country was no longer significant (p = .44). Hence, we decided to analyze all obtained effects. Finally, we are making available the raw data and R scripts to researchers interested in replicating our analyses, available here: https://osf.io/zhr8v/.

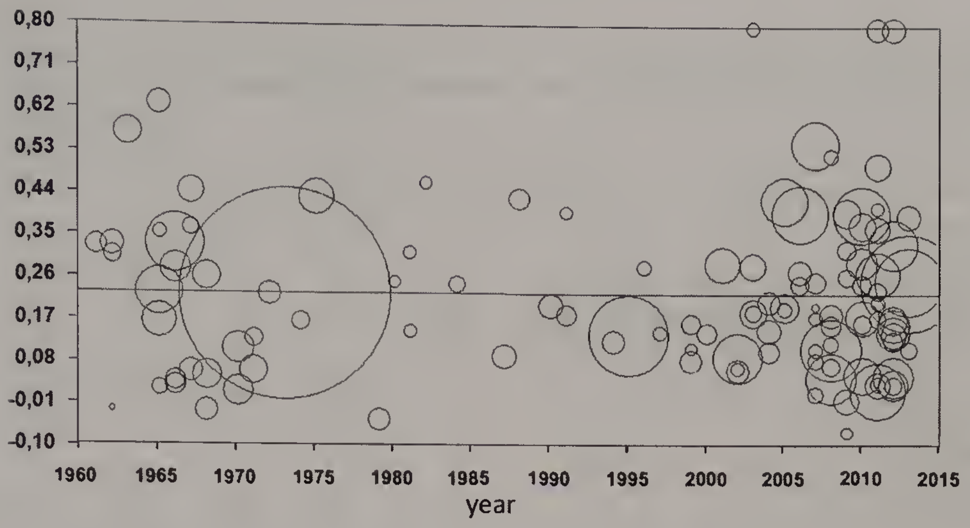


Figure 5. The relations between study year and effect size.

Finally, in an effort to provide a summary of estimated effects of the moderators, we conducted a meta-analysis analog of ANOVA[9] using the estimations obtained at the study level. Results of that analysis are presented in Table 6. As with our previously reported findings, the results of this analysis indicate that the observed effect was stable across time (similar in concurrent decades) but moderated by the type of creative test used. More specifically, verbal tests developed in the Guilford tradition (e.g., unusual uses or consequences tasks; see Guilford, 1967) resulted in more than two times higher correlations (r = .30) with academic achievement than did figural (e.g., TCT-DP, see Urban, 1991) tests (r = .14). Moreover, the use of standardized academic achievement tests resulted in higher correlations with creativity (r = .28) compared to the use of GPA (r = .19). In addition, the academic stage of middle school (r = .33) resulted in higher correlations between creativity and academic achievement compared to elementary schools (r = .23), high schools (r = .21), and universities (r = .17).

Finally, the results of our ANOVA analog analysis also indicated significant differences in the strength of creativity and academic achievement between continents (Q = 32.58, df = 5, p < .001). This finding, however, suggests that it is an artifact caused by the lack of control for differences in the characteristics of studies. Indeed, as our previous analysis indicates (see the online supplemental material Table S1), when properly controlling for between-country differences in creativity and academic achievement measurement, continent does not significantly influence the obtained the effect size.

Discussion

The goal of this meta-analysis was to clarify the somewhat mixed findings of previous research that has examined the relationship between creativity and academic achievement. More specifically, we endeavored to obtain a stable estimate of the direction and magnitude of the relationship between creativity and academic achievement. In addition, we had the aim of examining the influence of potential moderators on this observed relationship.

What Is the Relationship Between Creativity and Academic Achievement?

With respect to the relationship between creativity and academic achievement, our results indicate that there is a modest but significantly positive association (r = .22) in the studies we analyzed. Moreover, our analyses indicate that this relationship was not influenced by publication bias. These findings align with long-standing assertions of scholars who have described creativity and learning as representing interrelated phenomena (e.g., Beghetto, 2016a; Guilford, 1967; Piaget, 1962, 1981; Sawyer, 2012; Vygotsky, 1967/2004). The modest magnitude of this relationship (r = .22), however, raises questions as to why the observed association was so low. Indeed, this relationship only explains 5% of the variance in creativity and academic achievement. With so much unaccounted for variance, it is important to consider what factors might be influencing this relationship. The results of our moderator analysis help shed some light on this issue. In the sections that follow, we discuss the results of our moderator analysis and conclude with a brief discussion of strengths, limitations, and future directions for this line of research.

What Is the Influence of Different Types of Measures?

Conceptually speaking, one of the clearest factors that can influence the observed relationship between creativity and academic achievement is how the constructs are measured. Our results indicate that the relationship between creativity and academic achievement was significantly stronger when creativity was measured with tests, r = .23, 95% CI [.20, .26]—particularly verbal tests, r = .30, 95% CI [.25, .34]—compared to when it was measured using self-report scales, r = .12, 95% CI [.07, .17]. That test-based measures would have a stronger influence on the relationship between creativity and academic achievement is not sur-

[9] Although this analytic technique does not control for the associations and shared variance between moderators (and is therefore less robust than previously reported multilevel regression models), it provides results (i.e., effects in terms of averaged correlations), which tend to be easier for readers to interpret.

Table 4
Moderator Analysis: Education Stage

Effects	Estimate	*SE*	95% CI *LL*	95% CI *UL*	*p*
Fixed effects					
Intercept	.17	.04	.08	.25	<.001
Creativity measurement (0 = self-report, 1 = test)	.08	.03	.02	.14	.01
Academic achievement measurement (0 = test, 1 = GPA)	−.03	.02	−.07	.004	.08
Education stage (elementary = reference category)					
Middle school	.12	.05	.02	.21	.015
High school	.004	.04	−.08	.08	.93
College/university	−.04	.04	−.11	.04	.36
Random effects					
Within-study variance	.01	.001	.008	.01	<.001
Between-study variance	.02	.003	.011	.022	<.001

Note. Estimated on 112 studies and 628 correlations. CI = confidence interval; *LL* = lower limit; *UL* = upper limit; GPA = grade point average.

prising. Indeed, as we noted earlier, cognitive characteristics relevant to creative ability, such as the fluency, flexibility, and originality of thinking (Guilford, 1967); imagination (Jankowska & Karwowski, 2015); induction and deduction abilities (Weisberg, 2006); and the use of specific problem-solving strategies play a considerable role in the learning process (Chamot, Dale, O'Malley, & Spanos, 1992; Hmelo-Silver, 2004). As such, our results provide further evidence of the potentially positive role that creativity can play in the acquisition, consolidation, and processing of new knowledge—including school knowledge (Hennessey & Amabile, 1987).

We also found that obtained effect size differed depending on the type of academic achievement measure used. More specifically, when the criterion of achievement was GPA, the effect was significantly weaker, $r = .19$, 95% CI [.16, .22], compared to when achievement was measured using standardized achievement tests, $r = .28$, 95% CI [.22, .34]. This difference may be caused by various factors. It may reflect the lower reliability of school grades compared to standardized achievement tests (Elliott & Strenta, 1988). In a majority of the meta-analyzed studies (especially the early ones), data concerning the reliability of grades were not given, and therefore we were unable to estimate the corrected correlations.

It is also possible, however, that this difference has substantive meaning. One reason why the correlation between creativity and grades was lower than the correlation between creativity and more objective academic achievement tests (Organisation for Economic Co-operation and Development [OECD], 2014) is because the willingness to express one's creativity can be influenced by subtle environmental features of the classroom (Amabile, 1996; Beghetto

Table 5
Moderator Analysis: Figural vs. Verbal Creativity Tests

Effects	Estimate	*SE*	95% CI *LL*	95% CI *UL*	*p*
Fixed effects					
Intercept	−.040	.131	−.296	.216	.76
Year (grand centered)	.001	.001	−.001	.003	.36
Goal (0 = other, 1 = creativity × achievement)	.018	.033	−.048	.083	.59
Published? (0 = no, 1 = yes)	.004	.035	−.064	.073	.90
Academic achievement measurement (0 = test, 1 = GPA)	−.020	.020	−.060	.019	.31
Test type (figural = 0, verbal = 1)	.170	.017	.136	.203	<.001
School subjects (sport = reference category)					
Humanistic	.200	.124	−.043	.443	.11
Science	.189	.124	−.055	.432	.13
Overall	.164	.127	−.084	.413	.20
Creative abilities (other + general = reference)					
Fluency	−.038	.023	−.083	.007	.10
Flexibility	−.024	.025	−.073	.025	.35
Originality	−.029	.024	−.076	.018	.23
Elaboration	.018	.030	−.041	.077	.56
Random effects					
Within-study variance	.009	.001	.007	.011	<.001
Between-study variance	.017	.003	.011	.023	<.001

Note. Estimated on 90 studies and 617 correlations. CI = confidence interval; *LL* = lower limit; *UL* = upper limit; GPA = grade point average.

Table 6
Meta-Analysis Analog of ANOVA: Summary of Moderators

Moderator	k	N	r	95% CI	Heterogeneity (Q)[a]
Decade (Q = 8.83, df = 5, p = .12)					
1960–1969	19	5,378	.25***	[.18, .32]	126.91***
1970–1979	8	17,198	.17***	[.09, .26]	46.37***
1980–1989	9	1,121	.20**	[.09, .31]	29.56***
1990–1999	7	3,024	.15***	[.11, .19]	6.25
2000–2009	35	10,239	.20***	[.15, .26]	273.57***
2010–2015	42	21,711	.23***	[.18, .27]	377.42***
Region (Q = 32.58, df = 5, p < .001)					
Africa	2	539	.03	[−.06, .11]	.39
South America	1	141	.16	[−.01, .32]	NA
North America	61	30,299	.22***	[.18, .26]	427.24***
Australia	1	855	.32***	[.26, .38]	NA
Asia	14	3,852	.27***	[.16, .38]	155.63***
Europe	41	22,985	.20***	[.16, .24]	272.55***
Type of creative ability mode (Q = 26.94, df = 2, p < .001)					
Verbal	42	18,929	.30***	[.25, .34]	438.16***
Nonverbal	28	10,451	.14***	[.10, .18]	62.93***
Creativity test (Q = 10.44, df = 3, p = .02)					
Guilford	30	11,125	.26***	[.21, .31]	205.19***
TCT-DP	15	3,929	.18***	[.14, .21]	13.78
TTCT	22	3,746	.20***	[.15, .25]	54.66***
Other	25	16,306	.27***	[.21, .34]	394.70***
Academic achievement measure (Q = 6.27, df = 1, p = .01)					
GPA	73	35,341	.19***	[.16, .22]	412.66***
Achievement tests	31	11,328	.28***	[.22, .34]	322.04***
Education stage (Q = 16.44, df = 3, p = .001)					
Elementary	26	10,906	.23***	[.17, .29]	204.21***
Middle school	15	8,511	.33***	[.27, .39]	204.73***
High school	28	21,559	.21***	[.16, .26]	148.82***
College/university	42	11,602	.17***	[.12, .22]	287.67***

Note. A meta-analysis analog of analysis of variance (ANOVA) was used with a study as a unit of analysis. k = the number of studies included in the analysis; N = sample size. CI = confidence interval; NA = Not Applicable; TCT-DP = Test of Creative Thinking–Drawing Production; TTCT = Torrance Test of Creative Thinking; GPA = grade point average.
[a] df for Q statistic is the number of studies (k) – 1.
*** $p < .001$.

& Kaufman, 2014; Hennessey, 2010). Teachers who, for instance, prioritize students' ability to meet predetermined task expectations (over originality) when assessing students' work send subtle messages to students that originality is not necessary or perhaps not wanted (Beghetto, 2013). Consequently, students may learn that it is not worth the risk or effort to try to be creative in their responses. It is also possible that teachers may downgrade more original or unexpected responses. Indeed, there is evidence that teachers sometimes hold negative views about student behaviors associated with creativity (Gralewski & Karwowski, 2013, 2016; Karwowski, 2007, 2010; Scott, 1999; Westby & Dawson, 1995). Regardless of the reason, it is important to note that the observed relationship was still positive (albeit, somewhat modest).

Taken together, these findings help illustrate the importance of the types of measures used to assess creativity and academic achievement. Indeed, given the theoretical links between creativity and learning, one might expect a stronger correlation than what we found. With respect to creativity, the most popular measures tend to focus on divergent thinking (i.e., the ability to produce original ideas) and less on convergent thinking (i.e., the ability to meet task constraints) (see Barbot, Besancon, & Lubart, 2015, for an exception). As such, commonly used creativity tests often fail to represent broader conceptions of creativity (Baer, 2014; Cropley, 2006), which include a combination of originality and task constraints (Beghetto, Kaufman, & Baer, 2015; Plucker et al., 2004; Simonton, 2012). Consequently, such measures are a bit too narrow in what they measure. The same can be said for self-assessments of creativity.

Indeed, it may be the case that self-assessments also suffer from a form of "originality bias" (Beghetto, 2010; Runco & Acar, 2010) wherein they emphasize the more divergent aspects of creativity at the expense of the more convergent aspects of creativity. Given that academic measures tend to focus more on convergence (i.e., meeting task constraints, providing expected results), the use of overly narrow measures of creativity may result in systematically suppressed estimates of the observed relationship between creativity and academic achievement. At this point, such assertions are somewhat speculative and therefore warrant attention in future studies. As such, future research should focus on developing and testing measures of creativity that more adequately represent the creative combination of divergent and convergent thinking (see Barbot et al., 2015; Lubart & Besançon, in press). Doing so may help clarify whether there is a stronger empirical relationship between creativity and academic achievement than what is otherwise represented in more traditional measures.

With respect to academic achievement, future studies should also use more precise measures of academic achievement. In the case of GPA, it is frequently effort (Brookhart, 1997), progress (Nitko, 2001), or even the student's adjustment to the teacher's demands (Wortham, 2004) that are evaluated. Moreover, given that creative students sometimes approach learning tasks in unexpected and unorthodox ways (Beghetto, 2013, 2016a; Günçer & Oral, 1993; Karwowski & Jankowska, in press), their GPAs may be negatively influenced by failing to meet behavioral expectations (rather than a reflection of academic ability). Measures of academic achievement that more clearly focus on learning gains (rather than meeting teachers' expectations for obtaining those gains) might provide a more accurate assessment of student learning and thereby more accurately reflect the relationship between student creativity and academic achievement.

What Is the Effect of Education Stage?

Our results indicate that the influence of education stage on the relationship between creativity and academic achievement is similar across most stages, with the exception of middle school (r = .33). Why might this be the case? Classic (Torrance, 1968) and more contemporary (Krampen, 2012) analyses suggest that although there may be declines in creativity development in childhood, there seems to be rather systematic growth in creative ability from puberty onward (Claxton, Pannells, & Rhoads, 2005; Milgram & Hong, 1999). Even though there is some evidence of higher levels of creativity in elementary school students compared to middle school students (Yi, Hu, Plucker, & McWilliams, 2013), middle school students may, on average, experience a boost in creative ability. This assertion has a basis in developmental theory (Feldman, 2003) and in neuropsychology (Barbot & Tinio, 2014). The middle school years are, for instance, thought to be a key developmental period for thinking skills, which are then measured in students' skills assessment programs such as Programme for International Student Assessment (OECD, 2014). Although studies have demonstrated an increase in thinking skills starting in elementary school (Molnár, Greiff, & Csapó, 2013), the most pronounced development of these skills tends to be the middle school years (Csapó, 1997). This is not to say that middle school years are free from declines or creative suppression (Beghetto & Dilley, 2016), but prior work suggests that these years of development may serve as a key time of growth in creative abilities (Barbot, Lubart, & Besancon, 2016; Kleibeuker, De Dreu, & Crone, 2013). Such assertions, however, warrant further empirical exploration.

Our findings also indicate higher correlations in the middle school stage of education compared to high school and universities. This finding has less theoretical and empirical support than the observed difference between elementary and middle school. One possible explanation is that learning becomes increasingly more specialized at higher levels of education. The majority of the studies included in our meta-analysis used general rather than discipline-specific measures of creative potential, which tend to have lower levels of predictive validity when explaining more specialized academic achievement (see Baer, 2014, in press). The fact that we did not observe differences in the strength of the relationship between various dimensions of school functioning may also be an indication that domain-general measures of creative ability—which tended to be operationalized as a form of divergent thinking (i.e., fluency, flexibility, elaboration, and originality of thinking)—were not sensitive enough to provide differential estimations of academic achievement across disciplines.

Once again, these findings point to the importance of the sensitivity and scope of the measures used to assess creativity and academic achievement. Indeed, both creativity and learning researchers tend to be in agreement that creativity and learning are domain specific (Alexander, 1995; Baer, 2014, in press; Beghetto et al., 2015; Poitras & Lajoie, 2013). Future research should therefore use domain-specific measures to examine whether such measures influence the observed relationship between creativity and learning and whether there are potentially important differences across domains.

What Is the Influence of Time and Place?

Finally, we examined the potential influence of time (i.e., when the study was conducted) and place (i.e., what country or continent the study was conducted). Our findings indicate that the relationship between creativity and academic achievement was stable across time and place. This finding differs from the results of previous research, which have suggested that creativity may be declining over time (Kim, 2011) and that creativity is often conceptualized and experienced differently across cultures (Kaufman & Sternberg, 2006).

When interpreting these findings, it is important to point out that the analyses conducted here and in related studies (e.g., Kim, 2005) are cross-sectional. Without longitudinal data, it is difficult (if not impossible) to make any definitive claims about the relationship between creativity and academic achievement across time. Moreover, the studies we analyzed did not have the goal of providing direct comparisons across cultures, and as such, cultural differences that may influence creativity and academic achievement may not have been adequately assessed or represented in the studies we analyzed. Consequently, strong claims about the influence of time and culture are not appropriate until additional research is conducted, which focuses specifically on addressing the impact of time (measured longitudinally) and the impact of culture (using more direct cross-cultural comparisons). Our findings, however, do provide a starting point for researchers to examine whether and under what conditions the positive relationship between creativity and academic achievement is stable across time and place.

Strengths and Limitations of the Present Study

Strengths

A strong point of our meta-analysis is that it serves as the first study to provide a stable estimate of the relationship between creativity and academic achievement. Consequently, this study contributes much-needed clarification on this relationship. Another key strength is the scope of the study. More specifically, our results cover a wide range of temporal (1962–2015), territorial (studies from all over the world), and numerical (120 independent studies, 782 effects, and the total joint sample exceeding 52,000 participants) factors. In fact, this study represents one of the largest meta-analyses in the creativity literature to date. We also consider the analytic models applied (multilevel meta-analysis) to be an

advantage. Indeed, multilevel models enabled us to provide more robust estimations of the observed effects and the effects of key moderators.

Limitations

A disadvantage of this meta-analysis was the limited number of moderators we were able to include. There are several moderating factors (e.g., instructional approach, curriculum used, contextual influences of schools and classrooms, and measures of various individual differences, such as student and teacher beliefs) that may have shed additional light on factors that influence the relationship between creativity and academic achievement. Additional studies are therefore needed that take into account these additional individual and sociocultural factors.

The unavailability of relevant data at the level of individual studies was also a limitation (e.g., the reliability of academic achievement measures). The lack of these data prevented us from being able to make corrections to the obtained effects. Future researchers (and journal reviewers) are therefore well advised to report (and require the reporting of) relevant psychometric data on all measures so that such corrections can be made.

Perhaps the most severe limitation of this synthesis was our inability to properly control for a number of mediators and confounding variables at the level of individual studies. This is a limitation that plagues meta-analytic studies more generally. One way to help address this issue is for researchers to ensure that their studies include as many theoretically important predictors of academic achievement in one study as possible. In the case of creativity, this would include factors such as intelligence and personality (Chamorro-Premuzic & Furnham, 2008; Day, Hanson, Maltby, Proctor, & Wood, 2010), thinking styles (Zhang, 2004, 2010, 2012; Zhang & Sternberg, 2005), motivational factors (Bandura, 1997; Hill & Amabile, 1993; Karwowski, 2011, 2012, 2014; Kaufman & Beghetto, 2013), and contextual factors (Beghetto & Kaufman, 2014; Schacter, Thum, & Zifkin, 2006).

As already mentioned, longitudinal studies, using more precise measures, are particularly needed. Longitudinal studies, although costly in terms of time and resources, would pay out in the form of being able to provide needed insights into how creativity and academic achievement grow and develop over time. Such studies would also enable researchers to empirically test various proposed theoretical links between creativity and academic achievement (Beghetto, 2016a), including whether the relationship is best thought of as unidirectional (e.g., creativity → academic achievement; academic achievement → creativity) or reciprocal (e.g., creativity ←→ academic achievement).

A final limitation we feel important to highlight pertains to the possibility of a nonlinear relationship between creativity and academic achievement. Such a relationship cannot be fully captured in the types of data (correlation coefficients) and analyses used in this study. A nonlinear pattern should therefore not be ruled out. Indeed, there is evidence that such patterns exist between creativity and related constructs, such as creativity and intelligence (see Jauk, Benedek, Dunst, & Neubauer, 2013; Karwowski & Gralewski, 2013).[10] Consequently, subsequent work should explore possible nonlinear patterns in the relationship between creativity and academic achievement using analytic techniques such as segmented regression (Jauk et al., 2013) or a "necessary condition analysis" (Dul, 2016).

Concluding Thoughts

For more than six decades, the question of whether creativity and academic achievement are related has been a focus of theoretical and empirical work in educational psychology. This question has proven to be a thorny one, complicated by various types of measures and potentially intervening factors. Not surprisingly, the results of previous research have run the gamut from positively related, unrelated, and even negatively related. The upshot of a decade's worth of research on this question is that it provided numerous effects that we were able to analyze using robust meta-analytic techniques and thereby take an important step in the direction of addressing the longstanding question of whether creativity and academic achievement are related.

Indeed, prior to this study, the question of whether there is a relationship between creativity and academic achievement could best be answered with the equivocal response of, "It depends." Based on the findings from this meta-analysis, we can now more confidently respond, "Previous research has, on average, demonstrated a positive (albeit modest) relationship between creativity and academic achievement, which is significantly moderated by the types of measures used to assess creativity and academic achievement." This, of course, does not mean that the question is now closed. Rather, the results of the present study provide researchers with a baseline correlation that they can use in subsequent research for comparison and further exploration.

The next logical step is to continue to design studies that examine the stability of this estimate and more carefully examine what additional factors might influence this relationship. We have already pointed to several needed directions for future study. One of the most important future directions pertains to developing and examining the influence of more precise measures of creativity and academic achievement. Such work, however, is not purely empirical. Complementary theoretical work is also needed to help specify how and to what extent creativity and academic achievement are related phenomena. Educational psychologists can play a key role in this endeavor by working alongside creativity researchers to develop more detailed theoretical models that help specify the relationship between creativity and academic achievement and also help develop more sensitive measures that can test and further clarify these asserted relationships. Doing so will provide addi-

[10] The nonlinear relationship between creativity and cognitive abilities, such as intelligence, has been asserted by some of the earliest theorists (e.g., Guilford, 1967). Some theorists have posited a so-called threshold hypothesis (see Jauk et al., 2013; Karwowski & Gralewski, 2013; Preckel, Holling, & Wiese, 2006). This hypothesis asserts a positive relationship between creativity and intelligence only in the groups of individuals whose intelligence level is below an IQ of 120, whereas above this threshold, the correlation is expected to disappear or weaken significantly (Guilford, 1967). Consequently, the threshold hypothesis does not assume linear association but rather a curvilinear inverted J-shaped relationship between intelligence and creativity. Similar thresholds may exist in the relationship between creativity and academic achievement, such as high levels of academic achievement suppressing creativity (see Simonton, in press) or, conversely, high levels of creativity negatively influencing academic achievement (Kim, 2008). We thank an anonymous reviewer for highlighting this possibility.

tional insights into the longstanding question of how creativity and academic achievement are related.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Al-Dhobaiban, N. (2005). *Exploring the relationship between self-regulation and creativity* (Unpublished doctoral dissertation). Alabama: University of Alabama.

Alexander, P. A. (1995). Superimposing a situation-specific and domain-specific perspective on an account of self-regulated learning. *Educational Psychologist, 30,* 189–193. http://dx.doi.org/10.1207/s15326985ep3004_3

Alexander, P. A., Schallert, D. L., & Reynolds, R. E. (2009). What is learning anyway? A topographical perspective considered. *Educational Psychologist, 44,* 176–192. http://dx.doi.org/10.1080/00461520903029006

Amabile, T. M. (1996). *Creativity in context.* Boulder, CO: Westview.

*Anderson, H. E., Jr., White, W. F., & Stevens, J. C. (1969). Student creativity, intelligence, achievement, and teacher classroom behavior. *The Journal of Social Psychology, 78,* 99–107. http://dx.doi.org/10.1080/00224545.1969.9922344

*Anwar, M. N., Aness, M., Khizar, A., Naseer, M., & Muhammad, G. (2012). Relationship of creative thinking with the academic achievements of secondary school students. *International Interdisciplinary Journal of Education, 1,* 44–47.

*Arseniuk, R. (2007). *Różnice w zakresie twórczości pomiędzy 12–13-letnimi chłopcami z lekką niepełnosprawnością intelektualną a ich pełnosprawnymi intelektualnie rówieśnikami* [Differences in creativity between the 12–13 year old boys with mild intellectual disabilities and their non-disabled peers] (Unpublished master's thesis). Warsaw, Poland: The Maria Grzegorzewska University.

*Auzmendi, E., Villa, A., & Abedi, J. (1996). Reliability and validity of a newly constructed multiple-choice creativity instrument. *Creativity Research Journal, 9,* 89–95. http://dx.doi.org/10.1207/s15326934crj0901_8

*Awamleh, H., Al Farah, Y., & El-Zraigat, I. (2012). The level of creative abilities dimensions according to Torrance Formal Test (B) and their relationship with some variables (sex, age, GPA). *International Educational Studies, 5,* 138–148.

Baer, J. (1993). *Divergent thinking and creativity: A task-specific approach.* Hillsdale, NJ: Erlbaum.

Baer, J. (2014). *Creativity and divergent thinking: A task-specific approach.* New York, NY: Psychology Press.

Baer, J. (in press). Why you are probably more creative (and less creative) than you think. In M. Karwowski & J. C. Kaufman (Eds.), *The creative self.* New York, NY: Academic Press.

Baltzer, S. W. (1988). A validation study of a measure of musical creativity. *Journal of Research in Music Education, 36,* 232–249.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York, NY: Macmillan.

Barbot, B., Besancon, M., & Lubart, T. I. (2015). Creative potential in educational settings: Its nature, measure, and nurture. *Education 3–13, 43,* 371–381.

Barbot, B., Lubart, T. I., & Besançon, M. (2016). "Peaks, slumps, and bumps": Individual differences in the development of creativity in children and adolescents. *New Directions for Child and Adolescent Development, 2016,* 33–45. http://dx.doi.org/10.1002/cad.20152

Barbot, B., & Tinio, P. P. (2015). Where is the "g" in creativity? A specialization-differentiation hypothesis. *Frontiers in Human Neuroscience, 8,* 1041.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50,* 1088–1101. http://dx.doi.org/10.2307/2533446

Beghetto, R. A. (2007). Ideational code-switching: Walking the talk about supporting student creativity in the classroom. *Roeper Review: A Journal on Gifted Education, 29,* 265–270. http://dx.doi.org/10.1080/02783190709554421

Beghetto, R. A. (2010). Creativity in the classroom. In J. C. Kaufman & R. J. Sternberg (Eds.), *Cambridge handbook of creativity* (pp. 447–464). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511763205.027

Beghetto, R. A. (2013). *Killing ideas softly? The promise and perils of creativity in the classroom.* Charlotte, NC: IAP Information Age Publishing.

Beghetto, R. A. (2016a). Creative learning: A fresh look. *Journal of Cognitive Education and Psychology, 15,* 6–23. http://dx.doi.org/10.1891/1945-8959.15.1.6

Beghetto, R. A. (2016b). *Big wins, small steps: How to lead for and with creativity.* Thousand Oaks, CA: Corwin.

Beghetto, R. A., & Dilley, A. E. (2016). Creative aspirations or pipe dreams? Toward understanding creative mortification in children and adolescents. *New Directions for Child and Adolescent Development, 2016,* 85–95. http://dx.doi.org/10.1002/cad.20150

Beghetto, R. A., & Kaufman, J. C. (2007). Toward a broader conception of creativity: A case for "mini-c" creativity. *Psychology of Aesthetics, Creativity, and the Arts, 1,* 73–79. http://dx.doi.org/10.1037/1931-3896.1.2.73

Beghetto, R. A., & Kaufman, J. C. (2014). Classroom contexts for creativity. *High Ability Studies, 25,* 53–69. http://dx.doi.org/10.1080/13598139.2014.905247

Beghetto, R. A., Kaufman, J. C., & Baer, J. (2015). *Teaching for creativity in the common core classroom.* New York, NY: Teachers College Press.

*Bentley, J. C. (1966). Creativity and academic achievement. *The Journal of Educational Research, 59,* 269–272. http://dx.doi.org/10.1080/00220671.1966.10883353

*Bierly, P. E., III, Kolodinsky, R. W., & Charette, B. J. (2009). Understanding the complex relationship between creativity and ethical ideologies. *Journal of Business Ethics, 86,* 101–112. http://dx.doi.org/10.1007/s10551-008-9837-6

Biostat. (2008). Comprehensive Meta-Analysis Software (Version 2.2.064) [Computer software]. Englewood, NJ: Author.

Blumen-Pardo, S. (2002). Effects of a teacher training workshop on creativity, cognition, and school achievement in gifted and non-gifted second-grade students in Lima, Peru. *High Ability Studies, 13,* 47–58. http://dx.doi.org/10.1080/13598130220132307

*Bowers, J. (1969). Interactive effects of creativity and IQ on ninth-grade achievement. *Journal of Educational Measurement, 6,* 173–177. http://dx.doi.org/10.1111/j.1745-3984.1969.tb00675.x

Bringuier, J. C. (1980). *Conversations with Jean Piaget.* Chicago, IL: University of Chicago Press.

Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied Measurement in Education, 10,* 161–180. http://dx.doi.org/10.1207/s15324818ame1002_4

Carson, S. H., Peterson, J. B., & Higgins, D. M. (2005). Reliability, validity, and factor structure of the creative achievement questionnaire. *Creativity Research Journal, 17,* 37–50. http://dx.doi.org/10.1207/s15326934crj1701_4

*Chamorro-Premuzic, T. (2006). Creativity versus conscientiousness: Which is a better predictor of student performance? *Applied Cognitive Psychology, 20,* 521–531. http://dx.doi.org/10.1002/acp.1196

*Chamorro-Premuzic, T., & Furnham, A. (2003). Personality predicts academic performance: Evidence from two longitudinal university sam-

ples. *Journal of Research in Personality, 37,* 319–338. http://dx.doi.org/10.1016/S0092-6566(02)00578-0

Chamorro-Premuzic, T., & Furnham, A. (2008). Personality, intelligence and approaches to learning as predictors of academic performance. *Personality and Individual Differences, 44,* 1596–1603. http://dx.doi.org/10.1016/j.paid.2008.01.003

Chamot, A. U., Dale, M., O'Malley, J. M., & Spanos, G. A. (1992). Learning and problem solving strategies of ESL students. *Bilingual Research Journal, 16,* 1–28. http://dx.doi.org/10.1080/15235882.1992.10162635

*Cheung, C. K., Rudowicz, E., Yue, X., & Kwan, A. S. F. (2003). Creativity of university students: What is the impact of field and year of study? *The Journal of Creative Behavior, 37,* 42–63. http://dx.doi.org/10.1002/j.2162-6057.2003.tb00825.x

Cheung, M. W. L. (2014). Fixed- and random-effects meta-analytic structural equation modeling: Examples and analyses in R. *Behavior Research Methods, 46,* 29–40. http://dx.doi.org/10.3758/s13428-013-0361-y

Cheung, M. W. L. (2015). metaSEM: An R package for meta-analysis using structural equation modelling. *Frontiers in Psychology, 5,* 1521. http://dx.doi.org/10.3389/fpsyg.2014.01521

Cheung, C.-K., Roskams, T., & Fisher, D. (2006). Enhancement of creativity through a one-semester course in university. *Journal of Creative Behavior, 40,* 1–25.

Childs, J. H. (1978). *Television viewing, achievement, IQ and creativity* (Unpublished doctoral dissertation). Provo, UT: Brigham Young University.

*Chooi, W. T., Long, H. E., & Thompson, L. A. (2014). The Sternberg Triarchic Abilities Test (Level-H) is a measure of g. *Intelligence, 2,* 56–67. http://dx.doi.org/10.3390/jintelligence2030056

Cicirelli, V. G. (1967). Sibling constellation, creativity, IQ, and academic achievement. *Child Development, 38,* 481–490. http://dx.doi.org/10.2307/1127304

*Clapham, M. M. (2001). The effects of affect manipulation and information exposure on divergent thinking. *Creativity Research Journal, 13,* 335–350. http://dx.doi.org/10.1207/S15326934CRJ1334_11

*Clapham, M. M. (2004). The convergent validity of the Torrance Tests of Creative Thinking and creativity interest inventories. *Educational and Psychological Measurement, 64,* 828–841. http://dx.doi.org/10.1177/0013164404263883

Claxton, A. F., Pannells, T. C., & Rhoads, P. A. (2005). Developmental trends in the creativity of school-age children. *Creativity Research Journal, 17,* 327–335. http://dx.doi.org/10.1207/s15326934crj1704_4

*Cline, V. B., Richards, J. M., Jr., & Abe, C. (1962). The validity of a battery of creativity tests in a high school sample. *Educational and Psychological Measurement, 22,* 781–784. http://dx.doi.org/10.1177/001316446202200416

*Cline, V. B., Richards, J. M., Jr., & Needham, W. E. (1963). Creativity tests and achievement in high school science. *Journal of Applied Psychology, 47,* 184–189. http://dx.doi.org/10.1037/h0041145

Cropley, A. (2006). In praise of convergent thinking. *Creativity Research Journal, 18,* 391–404. http://dx.doi.org/10.1207/s15326934crj1803_13

Csapó, B. (1997). Development of inductive reasoning: Cross-sectional measurements in an educational context. *International Journal of Behavioral Development, 20,* 609–626. http://dx.doi.org/10.1080/016502597385081

*Dąbrowska, M. (2008). *Zdolności twórcze, rezultaty szkolne i typ motywacji a atrakcyjność interpersonalna. Badania poprzeczne w grupach dwunasto-, piętnasto- i osiemnastolatków* [Creative abilities, academic achievement, type of motivation and interpersonal attractiveness. Cross study in groups of twelve, fifteen and eighteen-year-olds] (Unpublished master's thesis). Warsaw, Poland: The Maria Grzegorzewska University.

*Dai, D. Y., Tan, X., Marathe, D., Valtcheva, A., Pruzek, R. M., & Shen, J. (2012). Influences of social and educational environments on creativity during adolescence: Does SES matter? *Creativity Research Journal, 24,* 191–199. http://dx.doi.org/10.1080/10400419.2012.677338

*Day, L., Hanson, K., Maltby, J., Proctor, C., & Wood, A. (2010). Hope uniquely predicts objective academic achievement above intelligence, personality, and previous academic achievement. *Journal of Research in Personality, 44,* 550–553. http://dx.doi.org/10.1016/j.jrp.2010.05.009

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35,* 13–21. http://dx.doi.org/10.1016/j.intell.2006.02.001

*DeBoer, D. L. (1964). *A study of the relationship of creativity to intelligence and achievement* (Unpublished doctoral dissertation). Evanston, IL: Northwestern University.

*Dhatrak, G., & Wanjari, S. (2011). A co-relational study of scientific attitude, creativity and scholastic achievement of secondary school students. *Indian Streams Research Journal, 1,* 1–4.

*Diakidoy, I. A. N., & Constantinou, C. P. (2001). Creativity in physics: Response fluency and task specificity. *Creativity Research Journal, 13,* 401–410. http://dx.doi.org/10.1207/S15326934CRJ1334_17

Di Domenico, S. I., & Fournier, M. A. (2015). Able, ready, and willing: Examining the additive and interactive effects of intelligence, conscientiousness, and autonomous motivation on undergraduate academic performance. *Learning and Individual Differences, 40,* 41–54.

Dobrołowicz, J. (2002). *Kreatywność uczniów a ich osiągnięcia szkolne* [Creativity of students and their academic achievement]. Kielce, Poland: Wydawnictwo Akademii Świętokrzyskiej/Swietokrzyska Academy Press.

*Dollinger, S. J. (2011). "Standardized minds" or individuality? Admissions tests and creativity revisited. *Psychology of Aesthetics, Creativity, and the Arts, 5,* 329–341. http://dx.doi.org/10.1037/a0023659

Donovan, S. M., & Bransford, J. D. (2005). *How students learn: History, mathematics, and science in the classroom.* Washington, DC: The National Academies Press.

*Dowd, R. J. (1966). *The relationship of creative thinking to college success* (Final report of Small Contract Project No. 5–8202). Washington, DC: U.S. Office of Education of the Department of Health, Education and Welfare.

Dul, J. (2016). Necessary Condition Analysis (NCA): Logic and methodology of "necessary but not sufficient" causality. *Organizational Research Methods, 19,* 10–52. http://dx.doi.org/10.1177/1094428115584005

Dunbar, K. (1997). How scientists think: Online creativity and conceptual change in science. In T. B. Ward, S. M. Smith, & S. Vaid (Eds.), *Conceptual structures and processes: Emergence, discovery, and change* (pp. 461–493). Washington, DC: APA Press.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463. http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x

*Dziedziewicz, D. (2013). *Creativity, academic achievement and learning styles.* Unpublished manuscript. Warsaw, Poland: The Maria Grzegorzewska University.

*Dziedziewicz, D., & Gajda, A. (2011). *Creativity vs. academic achievement in middle school.* Unpublished manuscript. Warsaw, Poland: The Maria Grzegorzewska University.

Dziedziewicz, D., & Karwowski, M. (2015). Development of children's creative visual imagination: A theoretical model and enhancement programmes. *Education 3–13, 43,* 382–392.

Edwards, M. P., & Tyler, L. E. (1965). Intelligence, creativity, and achievement in a nonselective public junior high school. *Journal of Educational Psychology, 56,* 96–99. http://dx.doi.org/10.1037/h0021899

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315,* 629–634. http://dx.doi.org/10.1136/bmj.315.7109.629

Eisenman, R., Platt, J. J., & Darbes, A. (1968). Creativity, intelligence, and achievement. *Psychological Reports, 22,* 749–754. http://dx.doi.org/10.2466/pr0.1968.22.3.749

Elliott, R., & Strenta, A. C. (1988). Effects of improving the reliability of the GPA on prediction generally and on comparative predictions for gender and race particularly. *Journal of Educational Measurement, 25,* 333–347. http://dx.doi.org/10.1111/j.1745-3984.1988.tb00312.x

Farzaneh, G., Hassan, A., Gholamreza, P., Mirsalaldin, E., Parviz, A., & Alireza, H. (2010). Relationship between creativity, grade point average, achievement motivation, age and entrepreneur ship among university students. *Australian Journal of Basic and Applied Sciences, 4,* 5372–5378.

*Feldhusen, J. F., Treffinger, D. J., Van Mondfrans, A. P., & Ferris, D. R. (1971). The relationship between academic grades and divergent thinking scores derived from four different methods of testing. *Journal of Experimental Education, 40,* 35–40. http://dx.doi.org/10.1080/00220973.1971.11011301

Feldman, D. (2003). Key issues in creativity and development. In K. Sawyer, V. John-Steiner, S. Moran, R. Sternberg, D. Feldman, J. Nakamura, & M. Csikszentmihalyi (Eds.), *Creativity and development* (pp. 219–242). New York, NY: Oxford University Press.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research, and applications.* Cambridge, MA: MIT Press.

Freund, P. A., & Holling, H. (2008). Creativity in the classroom: A multilevel analysis investigating the impact of creativity and reasoning ability on GPA. *Creativity Research Journal, 20,* 309–318. http://dx.doi.org/10.1080/10400410802278776

*Freund, P. A., Holling, H., & Preckel, F. (2007). A multivariate, multilevel analysis of the relationship between cognitive abilities and scholastic achievement. *Journal of Individual Differences, 28,* 188–197. http://dx.doi.org/10.1027/1614-0001.28.4.188

*Gajda, A. (2008). *Pudło czy strzał w dziesiątkę? Trafność nominacji nauczycielskich a rzeczywisty poziom zdolności twórczych uczniów* [Miss or the jackpot? Relevance of teachers' nominations and the actual level of creative abilities of students] (Unpublished master's thesis). Warsaw, Poland: The Maria Grzegorzewska University.

*Gajda, A. (2013). *Creativity vs. academic achievement in primary school* (Unpublished manuscript). Warsaw, Poland: The Maria Grzegorzewska University.

Gallagher, S. (1985). *A comparison of Dabrowski's concept of overexcitabilities with measures of creativity and school achievement in sixth-grade student* (Unpublished master's thesis). Tucson: University of Arizona.

Getzels, J. W., & Jackson, P. W. (1962). *Creativity and intelligence: Explorations with gifted students.* New York, NY: Wiley.

*Gluskinos, U. M. (1971). Criteria for student engineering creativity and their relationship to college grades. *Journal of Educational Measurement, 8,* 189–195. http://dx.doi.org/10.1111/j.1745-3984.1971.tb00924.x

*Gollmar, S. M. (2000). *An investigation of attention deficit/hyperactivity disorder, creativity, and cognitive style: Interaction and impact on school success* (Unpublished doctoral dissertation). Athens: University of Georgia.

Gralewski, J., & Karwowski, M. (2012). Creativity and school grades: A case from Poland. *Thinking Skills and Creativity, 7,* 198–208. http://dx.doi.org/10.1016/j.tsc.2012.03.002

*Gralewski, J., & Karwowski, M. (2013). Polite girls and creative boys? Students' gender moderates accuracy of teachers' ratings of creativity. *The Journal of Creative Behavior, 47,* 290–304. http://dx.doi.org/10.1002/jocb.36

Gralewski, J., & Karwowski, M. (2016). Are teachers' implicit theories of creativity related to the recognition of their students' creativity? *The Journal of Creative Behavior.* http://dx.doi.org/10.1002/jocb.140

*Graves, G. O., Ingersoll, R. W., & Evans, L. R. (1967). The creative medical student: A descriptive study. *The Journal of Creative Behavior, 1,* 371–382. http://dx.doi.org/10.1002/j.2162-6057.1967.tb00068.x

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts—Something beyond *g*: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology, 105,* 364–379. http://dx.doi.org/10.1037/a0031856

Grigorenko, E. L., Jarvin, L., Diffley, R., Goodyear, J., Shanahan, E. J., & Sternberg, R. J. (2009). Are SSATs and GPA enough? A theory-based approach to predicting academic success in secondary school. *Journal of Educational Psychology, 101,* 964–981. http://dx.doi.org/10.1037/a0015906

*Grzelak, J. (2003). *Postawa twórcza i odtwórcza a osiągnięcia szkolne młodzieży III klas szkół średnich* [Creative and reproductive attitude and academic achievements of young people of secondary school third graders] (Unpublished master's thesis). Warsaw, Poland: The Maria Grzegorzewska University.

*Guastello, S. J., Bzdawka, A., Guastello, D. D., & Rieke, M. L. (1992). Cognitive abilities and creative behaviors: CAB-5 and consequences. *The Journal of Creative Behavior, 26,* 260–267. http://dx.doi.org/10.1002/j.2162-6057.1992.tb01185.x

Guilford, J. P. (1950). *Creativity. American Psychologist, 5,* 444–454.

Guilford, J. P. (1967). *The nature of human intelligence.* New York, NY: McGraw-Hill.

Günçer, B., & Oral, G. (1993). Relationship between creativity and nonconformity to school discipline as perceived by teachers of Turkish elementary school children, by controlling for their grade and sex. *Journal of Instructional Psychology, 20,* 208–214.

Halliburton-Beatty, C. L., & Simms, J. B. (2013). *Meta-analyses of effect of creativity on achievement.* Retrieved from http://jacquelinesimms.weebly.com/uploads/1/0/7/0/10705748/halliburtonbeattysimms-fiinalproject-edd1007.pdf

*Hansenne, M., & Legrand, J. (2012). Creativity, emotional intelligence, and school performance in children. *International Journal of Educational Research, 53,* 264–268. http://dx.doi.org/10.1016/j.ijer.2012.03.015

Hasirci, D., & Demirkan, H. (2007). Understanding the effects of cognition in creative decision making: A creativity model for enhancing the design studio process. *Creativity Research Journal, 19,* 259–271. http://dx.doi.org/10.1080/10400410701397362

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London, UK: Routledge.

Hennessey, B. A. (2010). Intrinsic motivation and creativity in the classroom: Have we come full circle? In R. A. Beghetto & J. C. Kaufman (Eds.), *Nurturing creativity in the classroom* (pp. 329–361). New York, NY: Cambridge University Press.

Hennessey, B. A., & Amabile, T. M. (1987). *Creativity and learning: What research says to the teacher.* Washington, DC: National Education Association.

*Hetrick, S. H., Lilly, R. S., & Merrifield, P. R. (1968). Figural creativity, intelligence, and personality in children. *Multivariate Behavioral Research, 3,* 173–187. http://dx.doi.org/10.1207/s15327906mbr0302_3

Hill, K. G., & Amabile, T. M. (1993). A social psychological perspective on creativity: Intrinsic motivation and creativity in the classroom and workplace. In S. G. Isaksen, M. C. Murdock, R. L. Firestien, & D. J. Treffinger (Eds.), *Understanding and recognizing creativity: The emergence of a discipline* (pp. 400–431). Norwood, NJ: Ablex Publishing.

*Hirsh, J. B., & Peterson, J. B. (2008). Predicting creativity and academic success with a "fake-proof" measure of the Big Five. *Journal of Re-*

search in Personality, 42, 1323–1333. http://dx.doi.org/10.1016/j.jrp.2008.04.006

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review, 16,* 235–266. http://dx.doi.org/10.1023/B:EDPR.0000034022.16470.f3

Hodder, S. M. (1972). *A study of the relationship between creativity and student teaching achievement* (Unpublished master's thesis). St. John's: Memorial University of Newfoundland.

*Hunter, S. T., & Cushenbery, L. D. (2015). Is being a jerk necessary for originality? Examining the role of disagreeableness in the sharing and utilization of original ideas. *Journal of Business and Psychology.* Advance online publication. doi:10.1007/s10869-014-9386-1

*Ibrahim, B. (2012). *Exploring the relationships among creativity, engineering knowledge, and design team interaction on senior engineering design projects* (Unpublished doctoral dissertation). Fort Collins: Colorado State University.

*Jackson, B. T. (1982). *The relationship of cognitive ability to achievement, creativity, and motivation of academically talented black primary level students in region VI of the District of Columbia public schools* (Unpublished doctoral dissertation). Yellow Springs: The Union for Experimenting Colleges and Universities.

*Janes, S. K. P. (1988). *The relationship between creativity test scores and honours course selection, honours course attrition, and grades among high school freshmen* (Unpublished doctoral dissertation). College Station: Texas A&M University.

*Jankowska, D. M., Gajda, A., & Karwowski, M. (2015). *Creative imagination and academic achievement: A longitudinal investigation.* Unpublished manuscript.

Jankowska, D. M., & Karwowski, M. (2015). Measuring creative imagery abilities. *Frontiers in Psychology, 6,* 1591. http://dx.doi.org/10.3389/fpsyg.2015.01591

Jauk, E., Benedek, M., Dunst, B., & Neubauer, A. C. (2013). The relationship between intelligence and creativity: New support for the threshold hypothesis by means of empirical breakpoint detection. *Intelligence, 41,* 212–221.

Jauk, E., Benedek, M., & Neubauer, A. C. (2014). The road to creative achievement: A latent variable model of ability and personality predictors. *European Journal of Personality, 28,* 95–105. http://dx.doi.org/10.1002/per.1941

*Joesting, J. (1975). Relationship of two tests of creativity to freshman English grades, school activities, and number of absences for black college students. *Psychological Reports, 37,* 69–70. http://dx.doi.org/10.2466/pr0.1975.37.1.69

*Johns, G. A., & Morse, L. W. (1997). Divergent thinking as a function of time and prompting to "be creative" in undergraduates. *The Journal of Creative Behavior, 31,* 156–165. http://dx.doi.org/10.1002/j.2162-6057.1997.tb00788.x

Johnson, W., McGue, M., & Iacono, W. G. (2007). How parents influence school grades: Hints from a sample of adoptive and biological families. *Learning and Individual Differences, 17,* 201–219. http://dx.doi.org/10.1016/j.lindif.2007.04.004

Kaltsounis, B. (1974). Creative behavior and achievement in mathematics. *Psychological Reports, 34,* 886. http://dx.doi.org/10.2466/pr0.1974.34.3.886

*Kaltsounis, B., & Stephens, H. G. (1973). Arithmetic achievement and creativity: Explorations with elementary school children. *Perceptual and Motor Skills, 36,* 1160–1162. http://dx.doi.org/10.2466/pms.1973.36.3c.1160

*Kanderian, S. S. (1969). *Study of the relationship between school achievement and measures of intelligence and creativity for students in Iraq* (Unpublished doctoral dissertation). Los Angeles: University of Southern California.

Karnes, M. B., Zehrbach, R. R., Studley, W. M., & Wright, W. R. (1965). *Culturally disadvantaged children of higher potential: Intellectual functioning and educational implications.* Springfield, IL: Departament of Program Planning for the Gifted, Office of the Superintendent of Public Instruction.

*Karwowski, M. (2005). *Rezultaty w nauce i kreatywność uczniów różniących się zdolnościami w percepcji nauczycieli* [Results in learning and the creativity of pupils with different abilities in the perception of teachers]. *Annales, 18,* 93–107.

Karwowski, M. (2007). Teachers' nominations of students' creativity: Should we believe them? Are the nominations valid? *Social Sciences, 2,* 264–269.

Karwowski, M. (2010). Are creative students really welcome in the classrooms? Implicit theories of "good" and "creative" student' personality among Polish teachers. *Procedia: Social and Behavioral Sciences, 2,* 1233–1237. http://dx.doi.org/10.1016/j.sbspro.2010.03.179

Karwowski, M. (2011). It doesn't hurt to ask . . . but sometimes it hurts to believe: Polish students' creative self-efficacy and its predictors. *Psychology of Aesthetics, Creativity, and the Arts, 5,* 154–164. http://dx.doi.org/10.1037/a0021427

Karwowski, M. (2012). Did curiosity kill the cat? Relationship between trait curiosity, creative self-efficacy and creative personal identity. *Europe's Journal of Psychology, 8,* 547–558. http://dx.doi.org/10.5964/ejop.v8i4.513

*Karwowski, M. (2013a). *Creativity vs. academic achievement 1* (Unpublished manuscript). Warsaw, Poland: The Maria Grzegorzewska University.

*Karwowski, M. (2013b). *Creativity vs. academic achievement 2* (Unpublished manuscript). Warsaw, Poland: The Maria Grzegorzewska University.

Karwowski, M. (2014). Creative mindsets: Measurement, correlates, consequences. *Psychology of Aesthetics, Creativity, and the Arts, 8,* 62–70. http://dx.doi.org/10.1037/a0034898

Karwowski, M. (2015, July). *Creativity and education: An awkward association.* Paper presented at the ICIE Excellence & Innovation in Education Conference: The Creativity—Innovation Challenge, Kraków, Poland.

Karwowski, M., Ciak, R., & Grubek, M. (2009). *Patrzy, ale czy widzi? Trafność nominacji nauczycielskich w rozpoznaniu kreatywności uczniów* (pp. 241–254). [He looks, but does he see? Relevance of teachers nominations in recognizing students' creativity]. In M. Karwowski (Ed.), *Identyfikacja potencjału twórczego. Teoria, metodologia, diagnostyka* [Identification of creative potential. Theory, methodology, diagnostics] Warsaw, Poland: The Maria Grzegorzewska University.

Karwowski, M., & Dziedziewicz, D. (2012). *Test Umiejętności na Starcie Szkolnym* [Test of skills at the start of school]. Warsaw, Poland: Instytut Badań Edukacyjnych/Institute for Educational Research.

Karwowski, M., & Gralewski, J. (2013). Threshold hypothesis: Fact or artifact? *Thinking Skills and Creativity, 8,* 25–33. http://dx.doi.org/10.1016/j.tsc.2012.05.003

Karwowski, M., & Jankowska, D. M. (in press). Four faces of creativity at school. In R. A. Beghetto & J. C. Kaufman (Eds.), *Nurturing creativity in the classroom.* New York, NY: Cambridge University Press.

Karwowski, M., & Lebuda, I. (2015). The Big Five, The Huge Two and creative self-beliefs: A meta-analysis. *Psychology of Aesthetics, Creativity, and the Arts.* Advance online publication. http://dx.doi.org/10.1037/aca0000035

*Karwowski, M., Lebuda, I., & Wiśniewska, E. (2008–2009). Creative abilities and styles as predictors of school successes. *Gifted and Talented International, 23–24,* 119–127.

*Karwowski, M., & Szumski, G. (2014). *Academic achievement* (Unpublished manuscript). Warsaw, Poland: The Maria Grzegorzewska University.

Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The four C model of creativity. *Review of General Psychology, 13,* 1–12. http://dx.doi.org/10.1037/a0013688

Kaufman, J. C., & Beghetto, R. A. (2013). In praise of Clark Kent: Creative metacognition and the importance of teaching kids when (not) to be creative. *Roeper Review: A Journal on Gifted Education, 35,* 155–165. http://dx.doi.org/10.1080/02783193.2013.799413

Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). *Essentials of creativity assessment.* New York, NY: Wiley.

Kaufman, J. C., & Sternberg, R. J. (Eds.). (2006). *The international handbook of creativity.* New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511818240

Kim, K. H. (2005). Can only intelligent people be creative? A meta-analysis. *Journal of Secondary Gifted Education, 16,* 57–66.

Kim, K. H. (2008). Underachievement and creativity: Are gifted underachievers highly creative? *Creativity Research Journal, 20,* 234–242. http://dx.doi.org/10.1080/10400410802060232

Kim, K. H. (2011). The creativity crisis: The decrease in creative thinking scores on the Torrance Tests of Creative Thinking. *Creativity Research Journal, 23,* 285–295. http://dx.doi.org/10.1080/10400419.2011.627805

*Kim, K. H., & Michael, W. B. (1995). The relationship of creativity measures to school achievement and to preferred learning and thinking style in a sample of Korean high school students. *Educational and Psychological Measurement, 55,* 60–74. http://dx.doi.org/10.1177/0013164495055001006

Kleibeuker, S. W., De Dreu, C. K., & Crone, E. A. (2013). The development of creative cognition across adolescence: Distinct trajectories for insight and divergent thinking. *Developmental Science, 16,* 2–12. http://dx.doi.org/10.1111/j.1467-7687.2012.01176.x

*Klein, S., Skager, R., & Erlebacher, W. (1966). Measuring artistic creativity and flexibility. *Journal of Educational Measurement, 3,* 277–286. http://dx.doi.org/10.1111/j.1745-3984.1966.tb00890.x

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods, 2,* 61–76. http://dx.doi.org/10.1002/jrsm.35

*Kousoulas, F. (2010). The interplay of creative behavior, divergent thinking, and knowledge base in students' creative expression during learning activity. *Creativity Research Journal, 22,* 387–396. http://dx.doi.org/10.1080/10400419.2010.523404

Krampen, G. (2012). Cross-sequential results on creativity development in childhood within two different school systems: Divergent performances in Luxembourg versus German kindergarten and elementary school students. *Europe's Journal of Psychology, 8,* 423–448. http://dx.doi.org/10.5964/ejop.v8i3.468

Lebuda, I., Zabelina, D., & Karwowski, M. (2015). Mind full of ideas: A meta-analysis of the mindfulness-creativity link. *Personality and Individual Differences.* Advance online publication.

*Lovelace, J. B., & Hunter, S. T. (2013). Charismatic, ideological, and pragmatic leaders' influence on subordinate creative performance across the creative process. *Creativity Research Journal, 25,* 59–74. http://dx.doi.org/10.1080/10400419.2013.752228

Lubart, T., & Besançon, M. (in press). On the measurement and mismeasurement of creativity. In R. A. Beghetto & B. Sriraman (Eds.), *Creative contradictions in education: Cross-disciplinary paradoxes and perspectives.* Chad, Switzerland: Springer International.

Malekian, F., & Fathi, H. (2012). Planning and production of mathematic teaching courseware for school elementary second grade and evaluation of its effect on academic achievement (summation topic) and learner's creativity. *Procedia: Social and Behavioral Sciences, 46,* 780–784. http://dx.doi.org/10.1016/j.sbspro.2012.05.198

*Mann, E. L. (2009). The search for mathematical creativity: Identifying creative potential in middle school students. *Creativity Research Journal, 21,* 338–348. http://dx.doi.org/10.1080/10400410903297402

*Marjoribanks, K. (1976). Academic achievement, intelligence, and creativity: A regression surface analysis. *Multivariate Behavioral Research, 11,* 105–118. http://dx.doi.org/10.1207/s15327906mbr1101_7

Marsh, H. W., & Hau, K. T. (2004). Explaining paradoxical relations between academic self-concepts and achievements: Cross-cultural generalizability of the internal/external frame of reference predictions across 26 countries. *Journal of Educational Psychology, 96,* 56–67. http://dx.doi.org/10.1037/0022-0663.96.1.56

*Matthew, C. T., & Stemler, S. E. (2013). Assessing mental flexibility with a new word recognition test. *Personality and Individual Differences, 55,* 915–920. http://dx.doi.org/10.1016/j.paid.2013.07.464

Mayfield, B. (1979). Teacher perception of creativity, intelligence and achievement. *Gifted Child Quarterly, 23,* 812–817.

*McCabe, M. P. (1991). Influence of creativity and intelligence on academic performance. *The Journal of Creative Behavior, 25,* 116–122. http://dx.doi.org/10.1002/j.2162-6057.1991.tb01361.x

*Mednick, M. T. (1963). Research creativity in psychology graduate students. *Journal of Consulting Psychology, 27,* 265–266. http://dx.doi.org/10.1037/h0042429

*Mednick, M. T., & Andrews, F. M. (1967). Creative thinking and level of intelligence. *The Journal of Creative Behavior, 1,* 428–431. http://dx.doi.org/10.1002/j.2162-6057.1967.tb00074.x

Milgram, R. M., & Hong, E. (1999). Creative out-of-school activities in intellectually gifted adolescents as predictors of their life accomplishments in young adults: A longitudinal study. *Creativity Research Journal, 12,* 77–87. http://dx.doi.org/10.1207/s15326934crj1202_1

Miller, A. L., Lambert, A. D., & Speirs Neumeister, K. L. (2012). Parenting style, perfectionism, and creativity in high-ability and high-achieving young adults. *Journal for the Education of the Gifted, 35,* 344–365. http://dx.doi.org/10.1177/0162353212459257

*Mitrofanow, K. (2013). *Wyobraźnia twórcza vs osiągnięcia szkolne* [Creative imagination vs. academic achievement] (Unpublished master's thesis). Warsaw, Poland: The Maria Grzegorzewska University.

Molnár, G., Greiff, S., & Csapó, B. (2013). Inductive reasoning, domain specific and complex problem solving: Relations and development. *Thinking Skills and Creativity, 9,* 35–45. http://dx.doi.org/10.1016/j.tsc.2013.03.002

Muhich, D. (1972). Cross-validation of a behavioral model for predicting school success. *Educational and Psychological Measurement, 32,* 439–442. http://dx.doi.org/10.1177/001316447203200221

Mumford, M. D., Medeiros, K. E., & Partlow, P. J. (2012). Creative thinking: Processes, strategies, and knowledge. *The Journal of Creative Behavior, 46,* 30–47. http://dx.doi.org/10.1002/jocb.003

*Naderi, H., Abdullah, R., Aizan, H. T., Sharir, J., & Kumar, V. (2009). Creativity, age and gender as predictors of academic achievement among undergraduate students. *Journal of American Science, 5,* 101–112.

*Niaz, M., Núñez, G. S., & Pineda, I. R. (2000). Academic performance of high school students as a function of mental capacity, cognitive style, mobility-fixity dimension, and creativity. *The Journal of Creative Behavior, 34,* 18–29. http://dx.doi.org/10.1002/j.2162-6057.2000.tb01200.x

Nitko, A. J. (2001). *Educational assessment of students.* Des Moines, IA: Prentice Hall.

Organisation for Economic Co-operation and Development (OECD). (2014). *PISA 2012 Results in focus: What 15-year-olds know and what they can do with what they know.* Paris, France: OECD Publishing.

*Ofili, P. (2011). *Do multicultural experiences and biculturalism promote creativity in international students?* (Unpublished master's thesis). Oxford: University of Mississippi.

*Ohnmacht, F. W. (1966). Achievement, anxiety and creative thinking. *American Educational Research Journal, 3,* 131–138. http://dx.doi.org/10.3102/00028312003002131

*Olatoye, R. A., Akintunde, S. O., & Yakasai, M. I. (2010). Emotional intelligence, creativity and academic achievement of business administration students. *Electronic Journal of Research in Educational Psychology, 8,* 763–786.

*Orieux, J. A. (1989). *Correlates of creative ability and performance in high school students* (Unpublished doctoral dissertation). Edmonton: University of Alberta.

Owen, S. V., Feldhusen, J. F., & Thurston, J. R. (1970). Achievement prediction in nursing education with cognitive, attitudinal, and divergent thinking variables. *Psychological Reports, 26,* 867–870. http://dx.doi.org/10.2466/pr0.1970.26.3.867

*Palaniappan, A. K., & Persekutuan, W. (2008). Influence of intelligence on the relationship between creativity and academic achievement: A comparative study. *International Journal of Learning, 15,* 267–277.

Peterson, R. A., & Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology, 90,* 175–181.

*Petrulyte, A. (2011). Bendrojo ugdymo mokyklų, kuriose sustiprintas meninis ugdymas, mokinių kūrybiškumas ir mokymosi sėkmė [Creativity and academic achievement of students at school with art specialization]. *Ugdymo psichologija, 22,* 48–54.

Piaget, J. (1962). *Play, dreams and imitation in childhood.* New York, NY: Basic Books.

Piaget, J. (1981). Creativity. In J. M. Gallagher & K. Reid (Eds.), *The learning theory of Piaget and Inhelder* (pp. 221–229). Monterey, CA: Brook/Cole.

*Piórkowska, S. (2012). *łLatwiej, szybciej i z sukcesem—dynamika zachowań strategicznych w uczeniu się i ich związki ze zdolnościami twórczymi* [It's easier, faster and successfully—the dynamics of strategic behavior in learning and their relationships with creative skills] (Unpublished master's thesis). Warsaw, Poland: The Maria Grzegorzewska University.

*Pishghadam, R., Khodadady, E., & Zabihi, R. (2011). Learner creativity in foreign language achievement. *European Journal of Educational Studies, 3,* 465–472.

Plucker, J., Beghetto, R. A., & Dow, G. (2004). Why isn't creativity more important to educational psychologists? Potential, pitfalls, and future directions in creativity research. *Educational Psychologist, 39,* 83–96. http://dx.doi.org/10.1207/s15326985ep3902_1

Poitras, E. G., & Lajoie, S. P. (2013). A domain-specific account of self-regulated learning: The cognitive and metacognitive activities involved in learning through historical inquiry. *Metacognition and Learning, 8,* 213–234. http://dx.doi.org/10.1007/s11409-013-9104-9

*Popov, A. (1992). Creativity and reading comprehension. *The Journal of Creative Behavior, 26,* 206–212. http://dx.doi.org/10.1002/j.2162-6057.1992.tb01177.x

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135,* 322–338. http://dx.doi.org/10.1037/a0014996

*Porter, J. N. (1974). Race, socialization and mobility in educational and early occupational attainment. *American Sociological Review, 39,* 303–316. http://dx.doi.org/10.2307/2094291

Preckel, F., Holling, H., & Wiese, M. (2006). Relationship of intelligence and creativity in gifted and non-gifted students: An investigation of threshold theory. *Personality and Individual Differences, 40,* 159–170. http://dx.doi.org/10.1016/j.paid.2005.06.022

*Pretz, J. E., & Kaufman, J. C. (in press). Do traditional admission criteria reflect applicant creativity? *The Journal of Creative Behavior.* http://dx.doi.org/10.1002/jocb.120

Priest, T. (2006). Self-evaluation, creativity, and musical achievement. *Psychology of Music, 34,* 47–61. http://dx.doi.org/10.1177/0305735606059104

*Putwain, D. W., Kearsley, R., & Symes, W. (2012). Do creativity self-beliefs predict literacy achievement and motivation? *Learning and Individual Differences, 22,* 370–374. http://dx.doi.org/10.1016/j.lindif.2011.12.001

R Development Core Team. (2013). *Nlme: Linear and nonlinear mixed effects models* (R package Version 3.1–103). Vienna, Austria: R Foundation for Statistical Computing.

Richards, R. L., & Casey, M. B. (1975, March). *Creativity and academic motivation variables as predictors of achievement in a two-year college for educationally marginal students.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

*Rindermann, H., & Neubauer, A. C. (2004). Processing speed, intelligence, creativity, and school performance: Testing of causal hypotheses using structural equation models. *Intelligence, 32,* 573–589. http://dx.doi.org/10.1016/j.intell.2004.06.005

*Ritchie, S. J., Luciano, M., Hansell, N. K., Wright, M. J., & Bates, T. C. (2013). The relationship of reading ability to creativity: Positive, not negative associations. *Learning and Individual Differences, 26,* 171–176. http://dx.doi.org/10.1016/j.lindif.2013.02.009

*Rodriguez, E. (1980). *Impact of creative training on academic achievement and creative thinking skills concerning four ethnic-sex groups in the fourth grade* (Unpublished doctoral dissertation). Las Cruces: New Mexico State University.

*Roke, L., & Kālis, E. (2015). Is there a link between creativity and school grades? Research with 9th grade students. *International Journal of Psychology: A Biopsychosocial Approach, 16,* 7–22.

Rudowicz, E. (2003). Creativity and culture: A two way interaction. *Scandinavian Journal of Educational Research, 47,* 273–290. http://dx.doi.org/10.1080/00313830308602

Runco, M. A. (1991). The evaluative, valuative, and divergent thinking of children. *The Journal of Creative Behavior, 25,* 311–319. http://dx.doi.org/10.1002/j.2162-6057.1991.tb01143.x

Runco, M. A., & Acar, S. (2010). Do tests of divergent thinking have an experiential bias? *Psychology of Aesthetics, Creativity, and the Arts, 4,* 144–148. http://dx.doi.org/10.1037/a0018969

Runco, M. A., & Albert, R. S. (1985). The reliability and validity of ideational originality in the divergent thinking of academically gifted and non-gifted children. *Educational and Psychological Measurement, 45,* 483–501.

Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R., & Waters, S. D. (2009). Does socioeconomic status explain the relationship between admissions tests and post-secondary academic performance? *Psychological Bulletin, 135,* 1–22.

*Sak, U., & Maker, J. (2006). Developmental variation in children's creative mathematical thinking as a function of schooling, age, and knowledge. *Creativity Research Journal, 18,* 279–291. http://dx.doi.org/10.1207/s15326934crj1803_5

Sawyer, R. K. (2012). *Explaining creativity: The science of human innovation* (2nd ed.). New York, NY: Oxford University Press.

Schacter, J., Thum, Y. M., & Zifkin, D. (2006). How much does creative teaching enhance elementary school students' achievement? *The Journal of Creative Behavior, 40,* 47–72. http://dx.doi.org/10.1002/j.2162-6057.2006.tb01266.x

Scott, C. L. (1999). Teacher's biases toward creative children. *Creativity Research Journal, 12,* 321–328. http://dx.doi.org/10.1207/s15326934crj1204_10

*Sethi, N. (2012). A study of academic achievement in mathematics in relation to creativity of high school students. *Indian Streams Research Journal, 2.* http://dx.doi.org/10.9780/22307850, http://isrj.org/UploadedData/981.pdf

*Silvia, P. J. (2008). Creativity and intelligence revisited: A latent variable analysis of Wallach and Kogan (1965). *Creativity Research Journal, 20,* 34–39. http://dx.doi.org/10.1080/10400410701841807

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143,* 534–547. http://dx.doi.org/10.1037/a0033242

Simonton, D. K. (2012). Taking the US Patent Office creativity criteria seriously: A quantitative three-criterion definition and its implications. *Creativity Research Journal, 24,* 97–106.

Simonton, D. K. (in press). Big-C versus little-c Creativity: Definitions,

implications, and inherent educational contradictions. In R. A. Beghetto & B. Sriraman (Eds.), *Creative contradictions in education: Cross disciplinary perspectives and paradoxes. Chad,* Switzerland: Springer International.

*Simpson, N. D. (1998). *Relationships between the academic achievement and the intelligence, creativity, motivation, and gender role identity of gifted children* (Unpublished doctoral dissertation). Houston, TX: University of Houston.

*Skager, R. W., Klein, S. P., & Schultz, C. B. (1967). The prediction of academic and artistic achievement at a school of design. *Educational Measurement: Issues and Practice, 4,* 105–117. http://dx.doi.org/10.1111/j.1745-3984.1967.tb00577.x

Son, M. J. (2009). *A study of Korean students' creativity in science using structural equation modelling* (Unpublished doctoral dissertation). Tucson: University of Arizona.

*Stallings, W. M., & Gillmore, G. M. (1972, April). *Relationships between figural creativity and grades in a college of fine and applied arts.* Paper presented at the American Educational Research Association Annual Meeting, Chicago, IL.

*Stankiewicz, A. (2006). *Uczeń zdolny w oczach nauczyciela* [Gifted student in the teacher's opinion] (Unpublished master's thesis). Warsaw, Poland: The Maria Grzegorzewska University.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5,* 60–78. http://dx.doi.org/10.1002/jrsm.1095

Sternberg, R. J. (1998). Cognitive mechanisms in human creativity: Is variation blind or sighted? *The Journal of Creative Behavior, 32,* 159–176. http://dx.doi.org/10.1002/j.2162-6057.1998.tb00813.x

Sternberg, R. J. (1999). A propulsion model of types of creative contributions. *Review of General Psychology, 3,* 83–100. http://dx.doi.org/10.1037/1089-2680.3.2.83

*Sternberg, R. J. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence, 34,* 321–350. http://dx.doi.org/10.1016/j.intell.2006.01.002

*Sternberg, R. J. (2007). Increasing academic excellence and enhancing diversity are compatible goals. *Educational Policy, 22,* 487–514. http://dx.doi.org/10.1177/0895904807310037

Sternberg, R., & Lubart, T. (1999). The concept of creativity: Prospect and paradigm. In R. Sternberg (Ed.), *Handbook of creativity* (pp. 3–15). Cambridge, NY: Cambridge University Press.

*Stiles, N. C. H. (1982). *Correlates of dimensions of creativity obtained by education graduate students* (Unpublished doctoral dissertation). Tempe: Arizona State University.

*Szumski, G., & Karwowski, M. (2014). *Academic achievement* (Unpublished manuscript). Warsaw, Poland: The Maria Grzegorzewska University.

*Tan, M., Mourgues, C., Bolden, D. S., & Grigorenko, E. L. (2014). Making numbers come to life: Two scoring methods for creativity in Aurora's Cartoon Numbers. *The Journal of Creative Behavior, 48,* 25–43. http://dx.doi.org/10.1002/jocb.39

*Tatlah, I. A., Aslam, T. M., Ali, Z., & Iqbal, M. (2012). Role of intelligence and creativity in the academic achievement of students. *International Journal of Physical and Social Sciences, 2,* 1–10.

*Tibbetts, J. W. (1968). *Relationships of creativity to socioeconomic status, race, sex, IQ, age and grade-point average in an adolescent population* (Unpublished doctoral dissertation). Los Angeles: University of Southern California.

*Tkaczyk, E. (2009). *Kreatywnie to znaczy razem . . . Style rozwiązywania konfliktów a wybrane aspekty kreatywności uczniów gimnazjum* [Creatively means together . . . Style of conflict resolution and some aspects of creativity among secondary school students] (Unpublished master's thesis). Warsaw, Poland: The Maria Grzegorzewska University.

*Toll, A. R. (1985). *Relationship between selected New York: Guilford Press structure-of-intellect semantic transformation tests and a measure of reading comprehension among sixth-grade students* (Unpublished doctoral dissertation). Philadelphia, PA: Temple University.

Torrance, E. P. (1968). *Torrance tests of creative thinking.* Princeton, NJ: Personnel Press.

Unal, H., & Demir, I. (2009). Divergent thinking and mathematics achievement in Turkey: Findings from the programme for international student achievement (PISA-2003). *Procedia: Social and Behavioral Sciences, 1,* 1767–1770. http://dx.doi.org/10.1016/j.sbspro.2009.01.313

Urban, K. K. (1991). On the development of creativity in children. *Creativity Research Journal, 4,* 177–191. http://dx.doi.org/10.1080/10400419109534384

*Uszyńska-Jarmoc, J. (2005). Different types of thinking of seven-year-old children and their achievements in school. *Early Child Development and Care, 175,* 671–680. http://dx.doi.org/10.1080/0300443042000302627

Vartanian, O., Martindale, C., & Kwiatkowski, J. (2003). Creativity and inductive reasoning: The relationship between divergent thinking and performance on Wason's 246 task. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 56,* 641–655. http://dx.doi.org/10.1080/02724980244000567

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36,* 1–48. http://dx.doi.org/10.18637/jss.v036.i03

*Vijetha, P., & Jangaiah, C. (2010). Intelligence, creative thinking abilities and academic achievement of children with hearing impairment—a correlation study. *Journal of the All India Institute of Speech & Hearing, 29,* 262–268.

*Vock, M., Preckel, F., & Holling, H. (2011). Mental abilities and school achievement: A test of a mediation hypothesis. *Intelligence, 39,* 357–369. http://dx.doi.org/10.1016/j.intell.2011.06.006

Vygotsky, L. S. (2004). Imagination and creativity in childhood. *Journal of Russian and East European Psychology, 42,* 7–97. (Original work published 1967).

*Wade, S. (1968). Differences between intelligence and creativity: Some speculation on the role of environment. *The Journal of Creative Behavior, 2,* 97–101. http://dx.doi.org/10.1002/j.2162-6057.1968.tb00089.x

*Wagner, S. B. (1972). *Timed and untimed measures of creativity and their correlates* (Unpublished doctoral dissertation). Detroit, MI: Wayne State University.

*Walia, P. (2012). Achievement in relation to mathematical creativity of eighth grade students. *Indian Streams Research Journal, 2.* http://dx.doi.org/10.9780/22307850, http://isrj.org/UploadedData/808.pdf

*Wallace, C. E., & Russ, S. W. (2015). Pretend play, divergent thinking, and math achievement in girls: A longitudinal study. *Psychology of Aesthetics, Creativity, and the Arts, 9,* 296–305. http://dx.doi.org/10.1037/a0039006

Wallas, G. (1926). *The art of thought.* New York, NY: Harcourt, Brace, & World.

*Wang, A. Y. (2007). *Contexts of creative thinking: Teaching, learning, and creativity in Taiwan and the United States* (Unpublished doctoral dissertation). Claremont, CA: Claremont Graduate University.

*Wasil, E. (2008). *Kreatywność ucznia a jego oceny i pozycja w oczach nauczyciela* [Student creativity, assessment and position in the opinion of the teacher] (Unpublished master's thesis). Warsaw, Poland: The Maria Grzegorzewska University.

Weisberg, R. W. (2006). *Creativity: Understanding innovation in problem solving, science, invention, and the arts.* New York, NY: Wiley.

Westby, E. L., & Dawson, V. L. (1995). Creativity: Asset or burden in the classroom? *Creativity Research Journal, 8,* 1–10. http://dx.doi.org/10.1207/s15326934crj0801_1

*Westcott, R. S. (1983). *Self-generated humour as an alternative in identifying potentially gifted, talented and creative high school students: An exploratory study* (Unpublished doctoral dissertation). Athens: University of Georgia.

Wilson, D. B. (2014). Meta-analysis macros for SAS, SPSS, and Stata. Retrieved October, 18, 2014, from http://mason.gmu.edu/~dwilsonb/ma.html

*Wódz, K. (1981). *Twórcze uzdolnienia uczniów a ich osiągnięcia szkolne* [The creative abilities of students and their academic achievement]. *Psychologia Wychowawcza, 5,* 577–586.

Wortham, S. (2004). From good student to outcast: The emergence of a classroom identity. *Ethos, 32,* 164–187. http://dx.doi.org/10.1525/eth.2004.32.2.164

Yamamoto, K. (1964). A further analysis of the role of creative thinking in high school achievement. *The Journal of Psychology: Interdisciplinary and Applied, 58,* 277–283. http://dx.doi.org/10.1080/00223980.1964.9916745

*Yamamoto, K. (1967). Creativity and unpredictability in school achievement. *The Journal of Educational Research, 60,* 321–325. http://dx.doi.org/10.1080/00220671.1967.10883504

*Yamamoto, K., & Chimbidis, M. E. (1966). Achievement, intelligence and creative thinking in fifth-grade children: A correlational study. *Merrill-Palmer Quarterly, 12,* 233–241.

*Yeh, Y. C. (2004). Seventh graders' academic achievement, creativity, and ability to construct a cross-domain concept map—a brain function perspective. *The Journal of Creative Behavior, 38,* 125–144. http://dx.doi.org/10.1002/j.2162-6057.2004.tb01236.x

Yi, X., Hu, W., Plucker, J. A., & McWilliams, J. (2013). Is there a developmental slump in creativity in China? The relationship between organizational climate and creativity development in Chinese adolescents. *The Journal of Creative Behavior, 47,* 22–40. http://dx.doi.org/10.1002/jocb.21

Yorke-Viney, S. (2007). *An examination of the effectiveness of arts integration in education on student achievement, creativity and self-perception* (Unpublished doctoral dissertation). Scranton, PA: Marywood University.

*Zabelina, D. L., Condon, D., & Beeman, M. (2014). Do dimensional psychopathology measures relate to creative achievement or divergent thinking? *Frontiers in Psychology, 5,* 1029.

*Zhang, L. F. (2004). Revisiting the predictive power of thinking styles for academic performance. *The Journal of Psychology: Interdisciplinary and Applied, 138,* 351–370. http://dx.doi.org/10.3200/JRLP.138.4.351-370

Zhang, L. F. (2010). Do thinking styles contribute to metacognition beyond self-rated abilities? *Educational Psychology, 30,* 481–494. http://dx.doi.org/10.1080/01443411003659986

Zhang, L. F. (2012). Why schools should care about intellectual styles. *Journal of Cognitive Education and Psychology, 11,* 256–270. http://dx.doi.org/10.1891/1945-8959.11.3.256

Zhang, L. F., & Sternberg, R. J. (2005). A threefold model of intellectual styles. *Educational Psychology Review, 17,* 1–53. http://dx.doi.org/10.1007/s10648-005-1635-4

*Zielińska, J. (2009). *Twór (-) czy wzór? Osoby znaczące dla młodzieży o zróżnicowanym poziomie zdolności twórczych i nasileniu postawy twórczej* [Creative master? Individuals significant for young people with different levels of severity creative abilities and creative attitude] (Unpublished master's thesis) Warsaw, Poland: The Maria Grzegorzewska University.

Received September 11, 2015
Revision received April 4, 2016
Accepted April 13, 2016 ■

Instructions to Authors

Journal of Educational Psychology

www.apa.org/pubs/journals/edu

The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Manuscript preparation. Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (6th ed.). Manuscripts may be copyedited for bias-free language (see pp. 70–77 of the *Publication Manual*). Formatting instructions (all copy must be double-spaced) and instructions on the preparation of tables, figures, references, metrics, and abstracts appear in the *Manual*. For APA's Checklist for Manuscript Submission, see www.apa.org/pubs/journals/edu. **Abstract and keywords.** All manuscripts must include an abstract containing a maximum of 250 words typed on a separate page. After the abstract, please supply up to five keywords or brief phrases. **References.** References should be listed in alphabetical order. Each listed reference should be cited in text, and each text citation should be listed in the References. Basic formats are as follows:

Hughes, G., Desantis, A., & Waszak, F. (2013). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychologial Bulletin, 139,* 133–151. http://dx.doi.org/10.1037/a0028566

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.

Gill, M. J., & Sypher, B. D. (2009). Workplace incivility and organizational trust. In P. Lutgen-Sandvik & B. D. Sypher (Eds.), *Destructive organizational communication: Processes, consequences, and constructive ways of organizing* (pp. 53–73). New York, NY: Taylor & Francis.

Adequate description of participants is critical to the science and practice of educational psychology; this allows readers to assess the results, determine generalizability of findings, and make comparisons in replications, extensions, literature reviews, or secondary data analyses. Authors should see guidelines for sample–subject description in the *Manual.* Appropriate indexes of effect size or strength of relationship should be incorporated in the results section of the manuscript (see p. 34 of the *Manual*). Information that allows the reader to assess not only the significance but also the magnitude of the observed effects or relationships clarifies the importance of the findings. **Figures.** Graphics files are welcome if supplied in TIFF or EPS format. APA's policy on publication of color figures is available at http://www.apa.org/pubs/authors/instructions.aspx?item=6.

Publication policies. APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. APA policy regarding posting articles on the Internet may be found at www.apa.org/pubs/authors/posting.aspx. In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 8.13). As this is a primary journal that publishes original material only, APA policy prohibits publication of any manuscript or data that have already been published in whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data on which their article depends. In addition, APA Ethical Principles specify that "after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 8.14). Authors must have available their data throughout the editorial review process and for at least 5 years after the date of publication.

Masked review policy. The *Journal* has a masked review policy, which means that the identities of both authors and reviewers are masked. Every effort should be made by the authors to see that the manuscript itself contains no clues to their identities. Authors should never use first person (*I, my, we, our*) when referring to a study conducted by the author(s) or when doing so reveals the authors' identities, e.g., "in our previous work, Johnson et al., 1998 reported that . . .." Instead, references to the authors' work should be in third person, e.g., "Johnson et al. (1998) reported that . . .." The authors' institutional affiliations should also be masked in the manuscript. Authors submitting manuscripts are required to include in the cover letter the title of the manuscript along with all authors' names and institutional affiliations. However, the first page of the manuscript should omit the authors' names and affiliations, but should include the title of the manuscript and the date it is submitted. Responsibility for masking the manuscript rests with the authors; manuscripts will be returned to the author if not appropriately masked. If the manuscript is accepted, authors will be asked to make changes in wording so that the paper is no longer masked. Authors are required to state in writing that they have complied with APA ethical standards in the treatment of their sample, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained at www.apa.org/ethics/ or by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242. APA requires authors to reveal any possible conflict of interest in the conduct and reporting of research (e.g., financial interests in a test procedure, funding by pharmaceutical companies for drug research). Authors of accepted manuscripts will be required to transfer copyright to APA.

Permissions. Authors of accepted papers must obtain and provide to the editor on final acceptance all necessary permissions to reproduce in print and electronic form any copyrighted work, including test materials (or portions thereof), photographs, and other graphic images (including those used as stimuli in experiments). On advice of counsel, APA may decline to publish any image whose copyright status is unknown.

Supplemental materials. APA can place supplementary materials online, which will be available via the published article in the PsycARTICLES database. To submit such materials, please see www.apa.org/pubs/authors/supp-material.aspx for details. Authors of accepted papers will be asked to work with the editor and production staff to provide supplementary materials as appropriate.

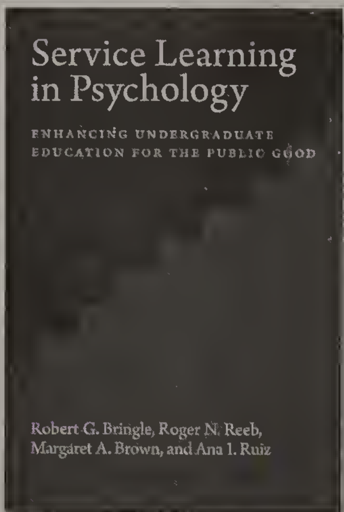
Submission. Authors should submit their manuscripts electronically via the Manuscript Submission Portal at www.apa.org/pubs/journals/edu/index.aspx (follow the link for submission under Instructions to Authors). General correspondence may be addressed to the incoming editorial office at AConley@apa.org.

American Psychological Association

SERVICE LEARNING IN PSYCHOLOGY

Enhancing Undergraduate Education for the Public Good

Robert G. Bringle, Roger N. Reeb, Margaret A. Brown, and Ana I. Ruiz



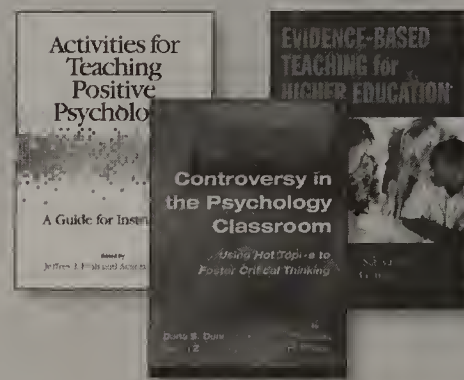
Service learning is a key component of undergraduate psychology, giving students—both majors and nonmajors—the unique opportunity to enhance their knowledge of psychology, develop critical thinking and communication skills, become more ethically and socially responsible, and evolve as scholars, professionals, and civic-minded citizens.

This book reviews the theory, research, and practice of service learning, showing how it can be integrated into the undergraduate psychology curriculum. Design strategies and sample projects are applied to introductory, major, and capstone courses that cover numerous subjects. Faculty development, departmental engagement, and the scholarship of teaching and learning are also examined. *2016. 256 pages. Hardcover.*
Available on Amazon Kindle®

List: $69.95 | APA Member/Affiliate: $49.95 | ISBN 978-1-4338-2079-3 | Item # 4316166

CONTENTS

ALSO OF INTEREST



Activities for Teaching Positive Psychology
A Guide for Instructors
Edited by Jeffrey J. Froh and Acacia C. Parks
2013. 173 pages. Paperback.
AVAILABLE ON AMAZON KINDLE®

List: $29.95 | APA Member/Affiliate: $29.95
ISBN 978-1-4338-1236-1 | Item # 4316148

Controversy in the Psychology Classroom
Using Hot Topics to Foster Critical Thinking
Edited by Dana S. Dunn, Regan A.R. Gurung, Karen Z. Naufel, and Janie H. Wilson
2013. 272 pages. Paperback.

List: $29.95 | APA Member/Affiliate: $24.95
ISBN 978-1-4338-1238-5 | Item # 4311017

Evidence-Based Teaching for Higher Education
Edited by Beth M. Schwartz and Regan A. R. Gurung
2012. 161 pages. Paperback.

List: $39.95 | APA Member/Affiliate: $39.95
ISBN 978-1-4338-1172-2 | Item # 4317288

PsycBOOKS®
Access to chapters from a variety of APA scholarly & professional books.

APA BOOKS ORDERING INFORMATION: 800-374-2721 • www.apa.org/pubs/books
In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502
In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972

AD3052